Resampling-based tuning of ordered model selection

Dissertation

Zur Erlangung des akademischen Grades Dr. rer. nat.

im Fach Mathematik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Diplom-Mathematiker Niklas Willrich,

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Elmar Kulke

Gutachter

1. Prof. Dr. Vladimir Spokoiny

2. Prof. Dr. Oleg Lepski

3. Prof. Dr. Enno Mammen

Tag der Verteidigung: 20.11.2015

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Berlin, den

Danksagungen

Zuerst möchte ich mich für die vielen hilfreichen Anregungen und die stete Diskussionsbereitschaft meines Betreuers Prof. Spokoiny bedanken. Außerdem möchte ich mich beim Weierstraß-Institut für Angewandte Analysis und Stochastik als Ganzem bedanken, das als Institut beste Bedingungen für die Erstellung dieser Arbeit geschaffen hat und mir die Arbeit an der Promotion auf einer Haushaltsstelle ermöglicht hat. Darüber hinaus möchte ich mich bei allen Kollegen der FG6 bedanken und natürlich speziell bei Andreas Andresen und Mayya Zhilova, die zeitgleich mit mir promoviert haben und mit denen man auch gerade in der Anfangszeit der Promotion immer wieder bei aufkommenden Fragen diskutieren konnte. Schlussendlich möchte ich mich noch bei meiner Frau bedanken, die mich über die ganze Promotionszeit hinweg liebevoll unterstützt hat.

Abstract

In this thesis, the *Smallest-Accepted* method is presented as a new Lepski-type method for ordered model selection. In a first step, the method is introduced and studied in the case of estimation problems with known noise variance. The main building blocks of the method are a comparison-based acceptance criterion relying on Monte-Carlo calibration of a set of critical values $\{\mathfrak{z}_{m,m^\circ}\}_{m,m^\circ\in\mathcal{M}}$ and the choice of the model as the smallest (in complexity) accepted model. The method can be used on a broad range of estimation problems like function estimation, estimation of linear functionals and inverse problems. General oracle results are presented for the method in the case of probabilistic loss and for a polynomial loss function. Applications of the method to specific estimation problems are studied.
In a next step, the method is extended to the case of an unknown possibly heteroscedastic noise structure. The Monte-Carlo calibration step is now replaced by a bootstrap-based calibration. A new set of critical values $\{\mathfrak{z}^\flat_{m,m^\circ}\}_{m,m^\circ\in\mathcal{M}}$ is introduced, which depends on the (random) observations. Theoretical properties of this bootstrap-based Smallest-Accepted method are then studied. It is shown for normal errors under typical assumptions, that the replacement of the Monte-Carlo step by bootstrapping in the Smallest-Accepted method is valid, if the underlying signal is Hölder-continuous with index $s > 1/4$ and $\log(n)\frac{p^2}{n}$ is small for a sample size n and a maximal model dimension p. In the proof of these results, some bounds of norms and traces for a class of random matrices based on Matrix-Bernstein inequalities are developed, which could be of independent theoretical interest.

Zusammenfassung

In dieser Arbeit wird die *Smallest-Accepted* Methode als neue Lepski-Typ Methode für Modellwahl im geordneten Fall eingeführt. In einem ersten Schritt wird die Methode vorgestellt und im Fall von Schätzproblemen mit bekannter Fehlervarianz untersucht. Die Hauptkomponenten der Methode sind ein Akzeptanzkriterium, basierend auf Modellvergleichen für die eine Familie von kritischen Werten $\{\mathfrak{z}_{m,m^\circ}\}_{m,m^\circ\in\mathcal{M}}$ mit einem Monte-Carlo-Ansatz kalibriert wird, und die Wahl des kleinsten (in Komplexität) akzeptierten Modells. Die Methode kann auf ein breites Spektrum von Schätzproblemen angewandt werden, wie zum Beispiel Funktionsschätzung, Schätzung eines linearen Funktionals oder Schätzung in inversen Problemen. Es werden allgemeine Orakelungleichungen für die Methode im Fall von probabilistischem Verlust und einer polynomialen Verlustfunktion gezeigt und Anwendungen der Methode in spezifischen Schätzproblemen werden untersucht.

In einem zweiten Schritt wird die Methode erweitert auf den Fall einer unbekannten, möglicherweise heteroskedastischen Fehlerstruktur. Die Monte-Carlo-Kalibrierung wird durch eine Bootstrap-basierte Kalibrierung ersetzt. Eine neue Familie kritischer Werte $\{\mathfrak{z}^\flat_{m,m^\circ}\}_{m,m^\circ\in\mathcal{M}}$ wird eingeführt, die von den (zufälligen) Beobachtungen abhängt. In Folge werden die theoretischen Eigenschaften dieser Bootstrap-basierten Smallest-Accepted Methode untersucht. Es wird gezeigt, dass unter typischen Annahmen unter normalverteilten Fehlern für ein zugrundeliegendes Signal mit Hölder-Stetigkeits-Index $s > 1/4$ und $\log(n)\frac{p^2}{n}$ klein, wobei n hier die Anzahl der Beobachtungen und p die maximale Modelldimension bezeichnet, die Anwendung der Bootstrap-Kalibrierung anstelle der Monte-Carlo-Kalibrierung theoretisch gerechtfertigt ist. Für den Beweis dieser Resultate werden einige Schranken für Normen und Spuren von einer Klasse von zufälligen Matrizen auf der Basis von Matrix-Bernstein-Ungleichungen entwickelt, die von eigenständigem theoretischen Interesse sein könnten.

Notation

Before we start with our exposition, we give a list of notations and typical naming conventions we will use in the following.

General:

$|\mathcal{M}|$ – denotes the cardinality of a set $\mathcal{M}$,

Ω – denotes a random set, usually of high probability which can change from line to line,

$\mathtt{C}$ – denotes some non-negative numerical constant, which can change from line to line too,

$\mathbb{1}(A)$ – denotes an indicator function of a set A,

$\mathcal{B}_p$ – denotes the σ-algebra of Borel measurable sets on $\mathbb{R}^p$,

Vectors & matrices:

$\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{Y}$ – bold-faced variable names will generally denote vector quantities,

$\mathbf{1}_m$ – indicates an m-dimensional identity matrix,

$\operatorname{diag}(\boldsymbol{u})$ – denotes a diagonal matrix with the coefficients of the vector $\boldsymbol{u}$ on the diagonal,

$\boldsymbol{u} \cdot \boldsymbol{w}$ – denotes a coordinate-wise product of two vectors: $\boldsymbol{u} \cdot \boldsymbol{w} = (u_i \cdot w_i)_{1 \le i \le n}$,

$A^\top$ – denotes the transpose of a matrix/vector A.

Norms:

$\|\boldsymbol{u}\|$ – denotes the standard Euclidean norm of a vector $\boldsymbol{u}$ in $\mathbb{R}^n$:

$$\|\boldsymbol{u}\| \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n} u_i^2}$$

$\|\boldsymbol{u}\|_\infty$ – denotes the sup-norm of a vector $\boldsymbol{u}$ in $\mathbb{R}^n$:

$$\|\boldsymbol{u}\|_\infty \stackrel{\text{def}}{=} \sup_{1\le i\le n} |u_i|.$$

$\|A\|_{\text{Fr}}$ – denotes the Frobenius norm of a matrix A:

$$\|A\|_{\text{Fr}} \stackrel{\text{def}}{=} \sqrt{\text{tr}(A^\top A)}.$$

$\|A\|_{\text{op}}$ – denotes the operator norm of a matrix $A \in \mathbb{R}^{p\times p}$:

$$\|A\|_{\text{op}} \stackrel{\text{def}}{=} \sqrt{\sup_{\gamma\in\mathbb{R}^p} \frac{\gamma^\top A^\top A\gamma}{\|\gamma\|^2}},$$

Statistical models and estimators:

$\boldsymbol{Y}$ – will generally denote a vector of observations,

Θ – denotes the parameter space of a model,

$L(\boldsymbol{\theta})$ – denotes the log-likelihood in parameter $\boldsymbol{\theta}$,

$\widetilde{\boldsymbol{\theta}}$ – will denote an estimator,

$\mathcal{K}(\cdot,\cdot)$ – denotes the Kullback-Leibler divergence between two probability distributions,

$\boldsymbol{\theta}^*$, $\boldsymbol{\theta}^*_m$, $\boldsymbol{f}^*$, ... – * will indicate a true parameter, or an optimal parameter, if the model is not well-specified,

n – usually denotes the sample size,

W – denotes a linear transformation of $\widetilde{\boldsymbol{\theta}}$,

ℓ – denotes a specific loss function for an estimation problem,

$\mathcal{R}$ – denotes a specific risk for an estimation problem,

Models:

m, m° – denote specific models,

$\mathcal{M}$ – denotes a set of models,

$\mathcal{M}^+(m)$ – denotes all models in $\mathcal{M}$, that are larger than m,

$\mathcal{M}^-(m)$ – denotes all models in $\mathcal{M}$, that are smaller than m,

$m_{\max}$ – denotes the largest model,

$m_{\min}$ – denotes the smallest model,

p – denotes the dimension of the largest model in a set of models $\mathcal{M}$,

$\Psi^\top$, $\Psi_m^\top$ – denote the design matrices for different linear models.

Characteristics of quadratic forms:

$\mathtt{p}_A$ – denotes the trace of a matrix $\mathrm{tr}(A)$, we will often call this an effective dimension,

λ_A – another short-hand for the operator norm of a matrix A,

Notation for SmA-method:

$\widetilde{\boldsymbol{\theta}}_m$ – denotes the estimator associated with the model m,

$\mathcal{S}_m$ – matrix defined by $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$,

$\mathcal{S}_{m,m^\circ}$ – denotes $\mathcal{S}_m - \mathcal{S}_{m^\circ}$,

W – denotes a linear transformation matrix applied to $\widetilde{\boldsymbol{\theta}}_m$,

$\widetilde{\boldsymbol{\phi}}_m$ – final estimator of our target for model m: $\widetilde{\boldsymbol{\phi}}_m = W\widetilde{\boldsymbol{\theta}} = \mathcal{K}_m \boldsymbol{Y}$.

$\mathcal{K}_m$ – matrix associated with $\widetilde{\boldsymbol{\phi}}_m$: $\mathcal{K}_m = W\mathcal{S}_m$,

$\mathcal{K}_{m,m^\circ}$ – denotes $\mathcal{K}_m - \mathcal{K}_{m^\circ}$,

$\mathbb{T}_{m,m^\circ}$ – norm of the difference of two estimators for different models : $\mathbb{T}_{m,m^\circ} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\|$,

$\boldsymbol{b}_{m,m^\circ}$ – bias component of an estimator difference,

$\boldsymbol{\xi}_{m,m^\circ}$ – stochastic component of an estimator difference,

$z_{m,m^\circ}(\cdot)$ – denotes the tail function for $\boldsymbol{\xi}_{m,m^\circ}$,

$z^+_{m,m^\circ}(\cdot)$ – denotes the tail-function with multiplicty-correction for $\boldsymbol{\xi}_{m,m^\circ}$,

$\mathfrak{z}_{m,m^\circ}$ – denotes the critical value for $\mathbb{T}_{m,m^\circ}$,

$\widehat{m}$ – denotes the model which is selected by the SmA-method.

Other:

$\xi^\flat, \mathfrak{p}^\flat$ – $\flat$ designates objects belonging to the Bootstrap-world.

Contents

Chapter 1

Introduction

Model selection is one of the key topics in mathematical statistics. How to choose between models of differing complexity is always a trade-off between overfitting the data by choosing a model, which has too many degrees of freedom and smoothing out underlying structure in the data by choosing a model which has too few degrees of freedom. This trade-off, which shows up in most methods as the classical bias-variance trade-off, is at the heart of every model selection method. Examples of current methods of model selection include penalized model selection [Barron et al., 1999], [Massart, 2007]), Lepski's method [Lepski, 1990], [Lepski, 1991], [Lepski, 1992], and risk hull minimization [Cavalier and Golubev, 2006]. We also mention cross-validation, which is especially popular with practitioners (see [Arlot and Celisse, 2010] for a survey).

Many of these methods allow their strongest theoretical results only for highly idealized situations, are very specific to one type of problem or have an unwieldy number of calibration constants whose choice is crucial to the applicability of the method.

The main contribution of this work is the introduction and theoretical study of a Lepski-type method of adaptive estimation that allows for a heteroscedastic noise structure and is applicable to a broad range of estimation problems. This so-called *Smallest-Accepted* method, in its most refined form, uses a bootstrap-based calibration procedure, which implicitly estimates the variance structure under some assumption of minimal smooth-

ness on the underlying signal and tunes its critical values to account for the dependencies between the different models.

Data-driven calibration of model selection is an active research topic. In [Spokoiny and Vial, 2009], a calibration procedure to get useful parameters for Lepski's method in the case of the estimation of a one-dimensional quantity of interest was introduced. One drawback, which is inherent in this method, is that exact knowledge of the noise level is crucial to its applicability. In [Arlot, 2009] the use of resampling methods for the choice of an optimal penalization was explored, following the framework of penalized model selection [Barron et al., 1999], [Massart, 2007]. Another approach of data-driven calibration in the face of an unknown error structure was proposed in [Arlot and Bach, 2009], [Birgé and Massart, 2007] using the concept of minimal penalties. These methods are based on ideas of penalized model selection and do not use a comparison-based model selection method like Lepski's method. The validity of a bootstrapping procedure for Lepski's method has recently been studied in [Chernozhukov et al., 2014] with new innovative technical tools. The authors develop results on honest adaptive confidence bands in a pointwise estimation setup in a non-Gaussian framework for the specific problem of Kernel density estimation.

The Smallest-Accepted method will allow for a heteroscedastic and unknown noise structure. It will be a Lepski-type method which calibrates its critical values by a propagation condition. The critical values are obtained by Monte-Carlo simulation in the case of known variance structure and by a bootstrap scheme in the case of an unknown possibly heteroscedastic variance structure. In the following, we will first review a number of model selection methods in Chapter 2. In Chapter 3, we then introduce and study the Smallest-Accepted method in the case of a known noise structure. The set-up covers linear regression and linear inverse problems, and equally applies to estimation of the whole parameter vector, its subvectors, as well as to the estimation of linear functionals. The proposed procedure and the theoretical study are also unified and do not distinguish between models and estimation problems. In the case of a linear inverse problem, the method is applicable to mildly and severely ill-posed problems without prior knowledge of the type and degree of ill-posedness; cf. [Tsybakov, 2000], [Cavalier et al.,

2002]. We will mainly consider linear models of the form $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in $\mathbb{R}^n$ for an unknown parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and a given design matrix $\Psi \in \mathbb{R}^{p \times n}$. We generally assume misspecification of the linear hypothesis, such that we can also treat nonparametric problems. One key assumption, which we will make throughout this thesis, is that the models we consider are totally ordered by their complexity. We require that for larger (in complexity) models, we have less bias, but more variability in the associated estimators. We will also generally suppose that for each model $m \in \mathcal{M}$, where $\mathcal{M}$ will denote our set of models, a linear estimator $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ is given, where $\mathcal{S}_m$ is a given $p \times n$ matrix. We will focus mainly on proving adaptivity of our model selection method. This means that we want to show that for our data-driven model selector $\widehat{m}$ the estimator $\widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ performs nearly as good as the optimal $\widetilde{\boldsymbol{\theta}}_{m^*}$ with m^* being the *oracle* model , which is unknown to us.

The basic idea of the *Smallest-Accepted* (SmA) method is based on a multiple-testing problem. The procedure can in fact be seen to consist of a family of pairwise tests: each model is tested against all larger ones and, if all tests pass, it is accepted. Finally the smallest accepted model is selected as our model estimator. The critical values for this multiple testing procedure are fixed using a so-called *propagation condition*. This condition basically demands that, if the variance of the estimator based on the model dominates its squared bias, then it should be accepted with high probability. To satisfy this condition the critical values are calibrated by Monte-Carlo simulation. This calibration step will adapt to the dependency structure of the test statistics and will usually give significantly smaller values for the multiplicity correction than a Bonferroni correction. Theorem 3.4.1 presents finite sample results on the behavior of the proposed selector $\widehat{m}$ and the corresponding estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. In particular, it describes a concentration set $\mathcal{M}^\circ$ for the selected model $\widehat{m}$ and states an oracle bound for the resulting estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. Usual rate results can easily be derived from these statements. Further results also address the size of the " payment for adaptation", which can be defined as the gap between the oracle and optimal adaptive bounds. Theorem 3.4.2 gives a general description of this quantity. We continue by specifying the results to special cases like prediction

of the whole function and the estimation of linear functionals. It appears that in some cases the obtained results yield sharp asymptotic bounds. In some other cases they lead to the usual log-price for adaptation. The results are mostly derived for probabilistic loss. However, in Theorem 3.5.1 of Section 3.5 we describe how the procedure and the results can be extended to the case of a polynomial loss function.

Chapter 4 extends the method and the theoretical study to the realistic case of an unknown heteroscedastic noise structure. The method automatically adjusts the parameters to the underlying possibly heterscedastic noise. The theoretical study becomes more challenging, because the critical values of the method are now random quantities too. It is shown for normal errors under typical assumptions, that the replacement of the Monte-Carlo step by bootstrapping in the Smallest-Accepted method is valid if the underlying signal is Hölder-continuous with index $s > 1/4$ and $\log(n)\frac{p^2}{n}$ is small for a sample size n and a maximal model dimension p. We also present some promising simulation results for the method on the typical problems of function estimation and the estimation of a first derivative. Further technical results used in the proofs are collected in the last chapter. Some bounds on different norms and traces of a specific class of random matrices, which are the key ingredients for the analysis of our bootstrap-based method, are given in Section 5.1. These bounds could also be of independent theoretical interest. The main results of this work are also going to appear in a more compact form in [Willrich and Spokoiny, 2015].

Chapter 2

Methods of model selection

In this chapter we explain what we understand by model selection and we give an overview of some of the standard methods of model selection. We will restrict our exposition to the case of ordered model selection, which means that we assume that we have an ordering of the complexity on the set of models $\mathcal{M}$. In the following, we will use $<$ to denote this total ordering. We assume that a less complex model will generally have more bias but less variability. We will specify this assumption more precisely in the following sections. Let us remark that we do not discuss the related topic of averaging of models (for this direction, see [Dalalyan and Salmon, 2012] and the contained references).

2.1 Framework of model selection

We are going to use two typical examples of estimation problems in the following. First we are going to define the sequence space framework for an observed vector $\boldsymbol{Y} \in I\!R^n$:

$$\boldsymbol{Y} = \boldsymbol{\theta}^* + \boldsymbol{\varepsilon},$$

with $n \in \mathbb{N} \cup \infty$, $\boldsymbol{\theta}^* \in I\!R^n$ being the true parameter and $\boldsymbol{\varepsilon} \in I\!R^n$ being a sequence of errors, which will often be assumed standard normal i.i.d. . The second one is the regression framework:

$$Y_i = f^*(x_i) + \varepsilon_i, 1 \leq i \leq n,$$

with $n \in \mathbb{N}$, $f^* : \mathbb{R}^d \to \mathbb{R}$ the true function and $\boldsymbol{\varepsilon} \in I\!R^n$ the error vector and $(x_i)_{1 \le i \le n} \subset \mathbb{R}^d$ the vector of design points. Here one can be interested in estimating different linear transformations of the true function or the true values at the design points. Typical examples include the estimation of the whole vector of values in the design points, the function at a specific point or some derivative of the function.

A statistical model is an assumption on the way the observed data is generated. The assumptions can be encoded by giving a family of likelihoods: $\{L_m\}_{m \in \mathcal{M}}$. We will assume a bit more general setting in that we also allow more general contrast functionals, e. g. likelihoods penalized by average curvature.

In the following, we will restrict ourselves to the situation of a finite-dimensional parametersets Θ_m and we mostly assume linear models for the data-generating process in the form:

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

for $\boldsymbol{\theta} \in I\!R^p$, $\boldsymbol{\varepsilon} \in I\!R^n$ and the design matrix $\Psi \in I\!R^{p \times n}$. In the following, we will assume a family of such models and we will not assume that the true data-generating process is an element of one of the models. We will often associate with the models some family of linear estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$. For example the least squares estimator in regression with normal errors or the maximum likelihood estimator in more general settings.

In this expository chapter, we will formulate most of the results for a sequence space setup. This approach makes it easier for us to emphasize similarities and connections between the different methods of model selection we present. Most of these methods work in more general setups. Another reason to look at a sequence space model is that one can show in many cases that more complicated setups can be transformed to give a sequence space model. One should however note that, in linear models, this is often based on the simultaneous diagonalization of the covariance matrix of the noise and the design, which is always possible for a homoscedastic noise structure, but need not be possible for heteroscedastic noise.

Aims of model selection

If we want to decide which member of a family of models to pick to model our data, we can have different objectives. Two main goals are discussed in the literature: One is to find a model selection procedure that chooses an estimator which makes the procedure adaptive. For this aim, we do not need that the data generating process is part of one of the models. The setting for adaptive estimation is often the following. Assume that we know that the true function belongs to a monotonously growing family of sets $(\Sigma_\beta)_{\beta \in B}$, where Σ_β could be a Sobolev ellipsoid of smoothness β for example and B some interval of possible smoothness parameters. If we knew the smallest $\beta' \in B$ such that $f^* \in \Sigma'_\beta$, we could use this information in the choice of an optimal estimator. An adaptive estimator tries to mimic (at the least in an asymptotic sense) the behavior of an optimal estimator constructed with the knowledge of β'. It is not always possible to attain the same asymptotic rates without the knowledge of β' as shown in [Lepski, 1990] for the white noise model with quadratic risk and in [Lepski, 1992] for more general setups. The difference in rates can be seen as a payment for adaptation.

A second possible aim of model selection is the identification of the true model. In the sense that, if we have a nested sequence of models, we want to choose the smallest model which contains the true value. The aim of model identification often necessitates a model selection method which tends to oversmooth, [Shao, 1997] and of course that the true value does not lie outside of all the models considered.

These two aims can be mutually exclusive, as shown in [Yang, 2005], if one demands adaptivity in a minimax-sense.

We will primarily discuss model selection methods striving for adaptivity. Hence we will often use the terms model selection and adaptive estimation interchangeably.

In the following, we assume given a risk $\mathcal{R}$, which is the expectation of some loss function ℓ — one can often think of quadratic loss and quadratic risk as a guiding example.

2.2 The oracle choice and the bias variance trade-off

We still focus on the *ordered case*. We assume given a set of models $\mathcal{M}$ and an associated set of estimators $\{\widetilde{\boldsymbol{\theta}}_m\}_{m\in\mathcal{M}}$. In some cases the set $\mathcal{M}$ of possible m choices can be countable and/or continuous and even unbounded. For simplicity of presentation, we assume that $\mathcal{M}$ is a finite set, $|\mathcal{M}|$ stands for its cardinality. We assume a risk function $\mathcal{R}$ and we define the oracle choice in the set of models for the risk $\mathcal{R}$ as:

$$m^* \stackrel{\text{def}}{=} \inf_{m\in\mathcal{M}} \mathcal{R}(\widetilde{\boldsymbol{\theta}}_m).$$

The oracle m^* is the model which gives the smallest risk for our estimation problem. Now we are going to explain how in the ordered case one can define the oracle by comparisons based on all the different estimators. We will give the argument in a sequence space model with normal independent errors and the quadratic risk. Consider:

$$\boldsymbol{Y} = \boldsymbol{\theta} + \sigma\boldsymbol{\varepsilon},$$

with $\boldsymbol{\theta}^* \in \mathbb{R}^n$ the true parameter, $\sigma > 0$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n)$ We use the projection estimator $\widetilde{\boldsymbol{\theta}}_m \stackrel{\text{def}}{=} (Y_i \mathbb{1}(i \le m))_{i\ge 1}$. This is the least squares estimator associated with the model

$$Y_i = \begin{cases} \theta_i + \varepsilon_i, & 1 \le i \le m, \\ \varepsilon_i, & i > m, \end{cases}$$

where $\boldsymbol{\theta} \in I\!R^m$. We also write $\boldsymbol{b}_m \stackrel{\text{def}}{=} (\theta_i^*)_{m+1\le i\le n}$ and $\boldsymbol{b}_{m^\circ,m} \stackrel{\text{def}}{=} (\theta_i^*)_{m^\circ+1\le i\le m}$ for $m > m^\circ$. The oracle can be written as:

$$m^* \stackrel{\text{def}}{=} \inf_{1\le m\le n} I\!E(\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2).$$

We have

$$\begin{aligned}
&\operatorname*{argmin}_{1\le m\le n} I\!E(\|(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2)\\
= &\operatorname*{argmin}_{1\le m\le n}(\mathrm{Var}(\widetilde{\boldsymbol{\theta}}_m) + \|\boldsymbol{b}_m\|^2)\\
= &\min_{1\le m\le n}\{\|\boldsymbol{b}_m\|^2 - \|\boldsymbol{b}_{m^\circ}\|^2 \le \mathrm{Var}(\widetilde{\boldsymbol{\theta}}_{m^\circ}) - \mathrm{Var}(\widetilde{\boldsymbol{\theta}}_m), \forall m^\circ > m\}\\
= &\min_{1\le m\le n}\{\|\boldsymbol{b}_{m,m^\circ}\|^2 \le \mathrm{Var}(\widetilde{\boldsymbol{\theta}}_{m^\circ} - \widetilde{\boldsymbol{\theta}}_m), \forall m^\circ > m\}.
\end{aligned}$$

So, in this special case we see that the oracle defined in the usual way is actually the value up to which the differences in squared bias are smaller than or equal to the variance for all comparisons with an $m^\circ \ge m$. This means that we can write:

$$m^* = \min_{1\le m\le n}\{\|\boldsymbol{b}_{m,m^\circ}\|^2 \le \mathrm{Var}(\widetilde{\boldsymbol{\theta}}_{m^\circ} - \widetilde{\boldsymbol{\theta}}_m), \forall m^\circ > m\} \qquad (2.2.1)$$

For the Smallest-Accepted method we are going to show our results with a definition of the oracle in the spirit of (2.2.1).

Next we are presenting the method of unbiased risk estimation.

2.3 Unbiased risk estimation

The basic idea of unbiased risk estimation (URE) is to replace the risk $\mathcal{R}$ one is trying to minimize by an unbiased estimator $\widetilde{\mathcal{R}}$ of it. One then chooses the minimizer of this unbiased estimator as the selected model:

$$\widetilde{m} \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{m\in\mathcal{M}} \widetilde{\mathcal{R}}(\widetilde{\boldsymbol{\theta}}_m).$$

In the sequence space model with known constant variance and quadratic risk and projection estimators as introduced above, the calculation is very simple (for more general models, we refer to [Stein, 1981]):

$$\begin{aligned}
&\ell(\boldsymbol{\theta})\\
= &\sum_{i=1}^{n}(\widetilde{\theta}_{m,i} - \theta_i^*)^2\\
= &\sum_{i=1}^{m}\varepsilon_i^2 + \sum_{i=m+1}^{n}\theta_i^{*2}
\end{aligned}$$

If one looks at the expectation of this object, one gets

$$\begin{aligned} & I\!E(\ell(\widetilde{\boldsymbol{\theta}}_m)) \\ &= \sigma^2 m + \sum_{i=m+1}^{n} \theta_i^{*2} \end{aligned}$$

As we do not know the bias term $\sum_{i=m+1}^{n} \theta_i^{*2}$, we need an unbiased estimator for it: We note that

$$\begin{aligned} & \sum_{i=1}^{n} (\widetilde{\theta}_{m,i} - Y_i)^2 - (n-m)\sigma^2 \\ = & \sum_{i=m+1}^{n} \left(\theta_i^{*2} + 2\varepsilon_i^2 \theta_i^* + {\varepsilon_i}^2 \right) - (n-m)\sigma^2 \end{aligned}$$

is such an unbiased estimator. Therefore the final estimator for the risk is

$$\widetilde{\mathcal{R}}(m) = \sum_{i=1}^{n} (\widetilde{\theta}_{m,i} - Y_i)^2 - (n-2m)\sigma^2$$

After stripping away parts which do not depend on m, we arrive at

$$\begin{aligned} \widetilde{m} = \operatorname*{argmin}_{1 \le m \le n} \widetilde{\mathcal{R}}(m) &= \operatorname*{argmin}_{1 \le m \le n} \sum_{i=1}^{n} (\widetilde{\theta}_{m,i} - Y_i)^2 + 2\sigma^2 m \\ = & \operatorname*{argmin}_{1 \le m \le n} \left(-\|\widetilde{\boldsymbol{\theta}}_m\|^2 + 2\sigma^2 m \right). \end{aligned}$$

We have arrived at the Akaike Information Criterion (AIC) [Akaike, 1974], respectively Mallow's C_p [Mallows, 1973], which are identical for this model. The argument only assures us of the unbiasedness of the $\widetilde{\mathcal{R}}(m)$. In the case of growing variance of the errors, we cannot be sure that the deviations $\widetilde{\mathcal{R}}(m) - \mathcal{R}(m)$ are of the same order of magnitude for all m. This leads to considering more general penalties. In the situation of heterogeneous noise with standard deviations $\sigma_1, \ldots, \sigma_n > 0$, one can write, following the same arguments,

$$\widetilde{m} = \operatorname*{argmin}_{1 \le m \le n} \widetilde{\mathcal{R}}(m) = \operatorname*{argmin}_{1 \le m \le n} \left(-\|\widetilde{\boldsymbol{\theta}}_m\|^2 + 2 \sum_{i=1}^{m} \sigma_i^2 \right).$$

One can see $-\sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{m} \sigma_i^2$ as the part corresponding to the risk we would get by plugging in $\boldsymbol{Y}$ for $\boldsymbol{\theta}^*$ after taking expectations and $\sum_{i=1}^{m} \sigma_i^2$

as a penalty term to correct for the bias. From this perspective, we can write

$$\widetilde{m} = \operatorname*{argmin}_{1\le m\le n} \widetilde{\mathcal{R}}(m) = \operatorname*{argmin}_{1\le m\le n} \left(-\|\widetilde{\boldsymbol{\theta}}_m\|^2 + \sum_{i=1}^{m} \sigma_i^2 + \mathtt{pen}(m) \right),$$

where $\mathtt{pen}(m) = \sum_{i=1}^{m} \sigma_i^2$. In the case where the variances $(\sigma_i)_{i\ge 1}$ of the errors grow in i, the URE method will meet problems. While it corrects for the bias of the risk estimate, it does not take into account possibly different orders of stochastic variation for different models. Optimality results in [Cavalier et al., 2002] show that this problems does not necessarily come up in purely asymptotic considerations (at least for moderately ill-posed problems). But [Cavalier and Golubev, 2006] points out that typical constants, which are hidden in an asymptotic setting, can be prohibitively large for typical inverse problems. The approach of penalized model selection gives a framework in which one can address this problem.

2.4 Penalized model selection

The basic idea of penalized model selection, in the sense presented in [Massart, 2007], is to take into account the deviation of the risk estimator $\widetilde{\mathcal{R}}(m)$ in a uniform manner and to add a model-dependent penalty term. In a Gaussian setup the tool for theoretical studies are Gaussian concentration inequalities. For the sequence space model with a projection estimator Theorem 4.2, [Massart, 2007], reads as follows with $\boldsymbol{\theta}_m^* \stackrel{\text{def}}{=} (\theta_i^*)_{1\le i\le m}$:

Theorem 2.4.1 (Thm. 4.2, [Massart, 2007])**.** *Let $(\mathtt{x}_m)_{m\in\mathcal{M}}$ be some family of positive numbers such that*

$$\sum_{m\in\mathcal{M}} \exp(-\mathtt{x}_m) = a < \infty. \tag{2.4.1}$$

Let $K > 1$ and assume that

$$\mathtt{pen}(m) \ge K\sigma^2(\sqrt{m} + \sqrt{2\mathtt{x}_m})^2.$$

Then, almost surely, there exists some minimizer $\widetilde{m}$ of the penalized least-

squares criterion:

$$-\|\widetilde{\boldsymbol{\theta}}_m\|^2 + \mathtt{pen}(m)$$

over $m \in \mathcal{M}$ *. Moreover, the corresponding penalized least-squares estimator* $\widetilde{\boldsymbol{\theta}}_{\widetilde{m}}$ *is unique and the following inequality holds*

$$I\!E_{\boldsymbol{\theta}^*}\left(\|\widetilde{\boldsymbol{\theta}}_{\widetilde{m}} - \boldsymbol{\theta}^*\|^2\right) \leq C(K)\left\{\inf_{m\in\mathcal{M}}\left(\|\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*\|^2 + \mathtt{pen}(m)\right) + (1+a)\sigma^2\right\}.$$

where $C(K)$ *depends only on* K *.*

If we take $K = 2$, we see that we get an enlarged-penalty version of AIC. The minimal choice of K to get sensible bounds is discussed in [Birgé and Massart, 2007] and it can be shown that the method fails in certain cases for $K < 1$. This can be used to calibrate a penalty in the case of an unknown error variance σ^2, by exploiting a phase transition in the behavior of the method around $K = 1$. We remark that the vector $(\mathtt{x}_m)_{m\in\mathcal{M}}$ has to be supplied by the user. The choice can be seen in some sense as a prior distribution on the set of models $\mathcal{M}$. In theoretical studies, the bound (2.4.1) is used for a Bonferroni correction. The dependency between different estimators is not taken into account for this model selection method.

2.5 Risk hull method

The risk hull method gives another approach to the choice of a penalty for model selection. Trying to find a way to better deal with the stochastic variation of an estimator of the risk, the method proposes a way to calibrate a penalty term based on Monte-Carlo simulations. It is built to deal with inverse problems, which characteristically exhibit a polynomial or even exponential increase in the variance of estimators of growing complexity.

The main conceptual contribution of this method is the introduction of the concept of a *risk hull*. We follow here the heuristic exposition in [Cavalier and Golubev, 2006]. We assume a sequence space model with growing variances. The risk of a projection estimator in this inverse problem setup is just $\sum_{i=m+1}^{n}\theta_i^2 + \sum_{i=1}^{m}\sigma_i^2\varepsilon_i^2$. Let us assume $\boldsymbol{\theta} \in I\!R^n$ known for

the moment. We look for a uniform upper bound on the stochastic error along all models

$$I\!E\left(\sup_{1\le m\le n}\left[\sum_{i=1}^{m}\sigma_i^2\varepsilon_i^2 - V(m)\right]\right) \le 0.$$

One can define

$$l(\theta,m) \stackrel{\text{def}}{=} \sum_{i=m+1}^{n}\theta_i^2 + V(m)$$

and call l a risk hull. The key inequality, which follows naturally from the definition of l, is

$$I\!E_{\boldsymbol{\theta}}\left(\|\widetilde{\boldsymbol{\theta}}_{\widetilde{m}} - \boldsymbol{\theta}\|^2\right) \le I\!E_{\boldsymbol{\theta}} l(\boldsymbol{\theta},\widetilde{m})$$

for any data-based model selector $\widetilde{m}$. This means that we can control the risk of any data-driven projection estimator, if we can control the risk hull. The important point is that l is non-random, which lets us avoid the problem of dealing with the typically rather complex dependence introduced in the estimators when one uses a data-driven model choice. Assuming polynomial growth of the σ_i^2, one can see that there exists $C>0$ such that for all $\alpha>0$:

$$l(\theta,m) = (1+\alpha)\left(\sum_{i=m+1}^{n}\theta_i^2 + \sum_{i=1}^{m}\sigma_i^2\right) + (1+\alpha)U_0(m) + \frac{C\sigma_1^2}{\alpha}$$

is a risk hull, where

$$U_0(m) \stackrel{\text{def}}{=} \inf\left\{t>0\colon I\!E\left(\eta_m 1\!\!1(\eta_m \ge t) \le \sigma_1^2\right)\right\}$$

with

$$\eta_m = \sum_{i=1}^{m}\sigma_i^2(\xi_i^2 - 1).$$

Replacing the unknown θ_i^2 by their unbiased estimates $Y_i^2 - \sigma_i^2$ leads to the following estimator:

$$m_{\text{rhm}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{m\ge 1}\left(-\sum_{i=1}^{m}y_i^2 + \sum_{i=1}^{m}\sigma_i^2 + \mathtt{pen}_{\text{rhm}}(m)\right)$$

with

$$\mathtt{pen}_{\mathrm{rhm}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{m} \sigma_i^2 + (1+\alpha)U_0(m),$$

We have the following oracle bound for the estimator:

Theorem 2.5.1 (Thm. 1, [Cavalier and Golubev, 2006])**.** *There exist constants* $C_* > 0$ *and* $\gamma_0 > 0$ *such that for all* $\gamma \in (0, \gamma_0]$ *and* $\alpha > 1$

$$I\!E\left(\|\widetilde{\boldsymbol{\theta}}_{m_{rhm}} - \theta\|^2\right) \le (1+\gamma)\inf_m R_{rhm}(\theta, m) + C_*\sigma_1^2\left(\frac{1}{\gamma^{4\beta+1}} + \frac{1}{\alpha - 1}\right),$$

where $R_{rhm}(\theta, m) = \sum_{k=m+1}^{n}\theta_k^2 + \sum_{k=1}^{m}\sigma_k^2 + (1+\alpha)U_0(m)$.

In [Cavalier and Golubev, 2006] it is proposed to approximate $U_0(\cdot)$ by Monte-Carlo simulation under the assumption, that one knows the noise variance $(\sigma_i^2)_{1\le i\le n}$.

2.6 Lepski's method

The methods we have presented before are based on the minimization of some criterion functional subject to a penalty. We will call this type of model selection method *penalty-based*. Another approach proposed by Lepski in [Lepski, 1990], [Lepski, 1991], [Lepski, 1992] is to compare all possible estimators in-between and choose a model on the basis of these comparisons, by selecting the "simplest estimator" which satisfies a certain acceptance criterion. We call methods which follow this general setup *Lepski-type* methods.

A Lepski-type method will mainly depend on two ingredients. The general form of an acceptance criterion and the specific critical values for the comparisons. Assume that we have an ordered family of models $\mathcal{M}$ and a family of estimators $(\widetilde{\boldsymbol{\theta}}_m)_{m\in\mathcal{M}}$. We now describe, what we will call the *classical* Lepski method. The procedure can be thought of as a sequential testing problem. For each $m \in \mathcal{M}$ we test the hypothesis

$$\widetilde{\boldsymbol{\theta}}_m = \widetilde{\boldsymbol{\theta}}_{m+1} = \widetilde{\boldsymbol{\theta}}_{m+2} = \ldots = \widetilde{\boldsymbol{\theta}}_{m_{\max}},$$

where $m_{\max}$ denotes the largest model. To this end, we choose critical values $(z_{m,m^\circ})_{m>m^\circ}$ and we sequentially check, starting with $m = m_{\max}$, if

$$\exists m^\circ > m; \|\widetilde{\boldsymbol{\theta}}_{m^\circ} - \widetilde{\boldsymbol{\theta}}_m\| > z_{m,m^\circ}.$$

If this is the case, we reject the hypothesis and choose the last model which passed the tests as our estimated model. Otherwise, we continue with the next smaller model. This means that

$$\widetilde{m} \stackrel{\text{def}}{=} \inf\left\{m \in \mathcal{M} \colon \forall m' \geq m^\circ \geq m, \|\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_{m^\circ}\| \leq z_{m^\circ,m'}\right\}.$$

The selected model therefore is the largest model, which is accepted and for which all larger models are accepted too. The method has been shown to be minimax-optimal for different problems of adaptive estimation [Lepski et al., 1997], [Lepski and Spokoiny, 1997]. One important challenge of this method is how to choose the critical values. Following [Spokoiny and Vial, 2009], we will give the idea of a data-driven method, which is designed for estimation of a function in a point. Here $\{\widetilde{\theta}_m\}_{m\in\mathcal{M}}$ is a family of one-dimensional estimators. The estimators are ordered by *increasing* complexity contrary to the convention in [Spokoiny and Vial, 2009]. Let θ^* denote the true value we are trying to estimate and $\theta^*_m \stackrel{\text{def}}{=} I\!E(\widetilde{\theta}_m)$ for $m \in \mathcal{M}$. We concentrate on the example of the quadratic risk. We decompose the estimators into a deterministic part and a stochastic part

$$\widetilde{\theta}_m = \theta^*_m + \xi_m$$

and assume that the stochastic part is Gaussian. Define

$$\mathtt{p}_m \stackrel{\text{def}}{=} I\!E\left(\xi_m^2\right).$$

The idea is to consider a sequence of bounds $z_{m^\circ} = z_{m^\circ,m'}$ for all $m^\circ > m' \geq m$, such that

$$I\!E_0\left((\widehat{\theta}_m - \widetilde{\theta}_m)^2\right) \leq \alpha \mathtt{p}_m,$$

where $\widehat{\theta}_m \stackrel{\text{def}}{=} \widetilde{\theta}_{\max\{\widetilde{m},m\}}$. So $\widehat{\theta}_m$ is an estimator which goes at least to the m-model and discards all models of lower complexity. $I\!E_0$ is the expectation

if we have no signal, i. e. $\theta^* = 0$. Under knowledge of the noise level one can approximate these bounds by Monte-Carlo simulation. This *propagation* condition basically means that we control the risk under the assumption of no noise, in the case where we do not stop before m, by the α-fraction of the effective dimension $\mathtt{p}_m$. This condition ensures that the effect of stopping too late is controlled, at least in the case where the bias is negligible. We assume the *oracle* choice m^* to be characterized by the bias-variance trade-off:

$$\max_{m^* \le m} b_m \le \beta v_{m^*}^{1/2} \tag{2.6.1}$$

with $b_m = |\theta_m^* - \theta^*|$ the bias of the model $m \in \mathcal{M}$ and $\beta \ge 0$. Under some technical conditions, we have the following result on the closeness of the oracle $\widetilde{\theta}_{m^*}$ and the estimator $\widehat{\theta}$ based on the critical values z_{m°:

Theorem 2.6.1 (Thm. 3.6, Thm. 3.8, [Spokoiny and Vial, 2009])**.** *For m^* satisfying* (2.6.1) *for some $\beta \ge 0$:*

$$I\!E\left(\mathtt{p}_{m^*}^{-1}(\widetilde{\theta}_{m^*} - \widehat{\theta})^2\right) \le \sqrt{\alpha}\mathtt{C}(\beta) + 2z(m^*),$$

where $\mathtt{C}(\beta)$ is a constant depending on β.

2.7 Tuning of model selection methods

The tuning step of the risk hull method and the calibration approach from [Spokoiny and Vial, 2009] are both based on the knowledge of the noise structure. In Chapter 3, we also first introduce the Smallest-Accepted method in such a framework. In comparison to the two methods we presented, our method can be used for a broader class of different estimation problems. It is not specific to a sequence space setup or estimation of a function in a point. One shortcoming of the method for known variance will still be its dependence on exact knowledge of the noise structure. Therefore, as we pointed out in the introduction, in Chapter 4, the Monte-Carlo step will be replaced by a bootstrap step. For model selection in the face of an unknown homoscedastic noise level, we first mention [Arlot and Bach, 2009], [Birgé

and Massart, 2007], where the noise level is estimated based on the penalized model selection framework and the existence of a minimal penalty level at which there is a qualitative change in behaviour of the model selection method. Following the notation of Theorem 2.4.1, the basic idea is that one varies $K' > 0$ in a penalty of the form $\texttt{pen}(m) = K'(\sqrt{m} + \sqrt{2\texttt{x}_m})^2$. One repeatedly applies the procedure for varying K' and as one can observe a phase transition in behavior at $K' = \sigma^2$, the location of the phase transition gives an estimate of the unknown noise level. We also cite [Bauer and Reiss, 2008] for a Lepski-type method which does not depend on knowledge of the homogeneous noise level and shows quasi-optimality under the assumption, of a nice prior distribution on the possible true functions.

In the case of heterogeneous noise, there has been work by [Arlot, 2009], where resampling of a penalization is used to get optimality results. The method proposed in the paper can be seen as a generalization of cross-validation to more general resampling schemes. We also mention [Chernozhukov et al., 2014] where the validity of a bootstrapping procedure for Lepski's method has recently been studied in a non-Gaussian situation. The authors develop results on honest adaptive confidence bands in a pointwise estimation setup for the specific problem of Kernel density estimation. The calibration of a Lepski-type method in a general regression framework with unknown heteroscedastic noise has, to the author's knowledge, not been treated so far in the literature. We now turn to the introduction of the method.

Chapter 3

Smallest-Accepted method with known noise variance

In the following chapter, we will introduce the SmA method for the case of a known noise variance. First, we present the basic framework for the method and which kinds of statistical problems we are considering in Section 3.1, then we are going to present the algorithm for model selection in Section 3.2 and the algorithm for calibration of the critical values in Section 3.3. Finally we are going to study the theoretical properties of the method in the last sections.

3.1 Notation and setting

Our point of departure is the following linear Gaussian model:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \;\text{ i.i.d. }, \qquad i = 1, \ldots, n, \quad (3.1.1)$$

with given design $\Psi_1, \ldots, \Psi_n \in I\!R^n$. We also write this equation in the vector form $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \in I\!R^n$ with the design matrix $\Psi \in I\!R^{n \times n}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n)$. Below we assume a deterministic design, otherwise one can understand the results conditioned on the design realization.

In what follows, we allow the model (3.1.1) to be misspecified. We mainly assume that the observations Y_i are independent and define the response

vector $\boldsymbol{f}^* = I\!E\boldsymbol{Y}$ with entries f_i^*. Such a model can be written as

$$Y_i = f_i^* + \varepsilon_i, \quad 1 \le i \le n. \tag{3.1.2}$$

In this chapter, we assume the noise distribution to be known. The main oracle results of Theorem 3.4.1 below do not require any further conditions on the noise. Some upper bounds on the quantities $\bar{\mathfrak{z}}_{m^*}$ entering in the oracle bounds are established under i.i.d. Gaussian noise, but could be extended to subgaussian heterogeneous noise under moment conditions.

For the linear model (3.1.2), we can write:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in I\!R^n} I\!E\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 = \left(\Psi\Psi^\top\right)^{-1}\Psi \boldsymbol{f}^*.$$

We also define $\mathcal{S}$ by $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$. As usual, we use a pseudo-inverse if the matrix $\Psi\Psi^\top$ is not invertible. The choice of n as a parameter dimension gives us a bias-free *linear* model for the signal $\boldsymbol{f}^*$. Let $\mathcal{M}$ designate a set of models and below we assume a family $\{\widetilde{\boldsymbol{\theta}}_m\}_{m \in \mathcal{M}}$ of linear estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ of $\boldsymbol{\theta}^*$ and define $\boldsymbol{\theta}_m^* \stackrel{\text{def}}{=} I\!E(\widetilde{\boldsymbol{\theta}}_m)$ for $m \in \mathcal{M}$. Typical examples include projection estimation onto an m-dimensional subspace or regularized estimation with a regularization parameter α_m, penalized estimators with a quadratic penalty function, etc. In the case of projection estimation onto an orthogonal basis, we will abuse notation slightly and write m for the model dimension too in cases where this makes sense (like projection estimation). To include specific problems like subvector/functional estimation and linear inverse problems, we also introduce a weighting matrix $W \in I\!R^{q \times p}$ for some fixed $q \ge 1$ and define the quadratic loss and risk weighted by this matrix W:

$$\ell_m \stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2,$$

$$\mathcal{R}_m \stackrel{\text{def}}{=} I\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

We are going to define the *probabilistic* loss for $K > 0$:

$$\mathbb{1}(\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 \ge K)$$

The associated risk is just

$$I\!P\left(\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 \ge K\right).$$

Of course, the loss and the risk depend on the specific choice of W. We do not indicate this dependence explicitly, but it is important to keep in mind the role of W in the definition of the losses.

Typical examples for a choice of W include:

Estimation of the whole parameter vector $\boldsymbol{\theta}^*$: We take W to be the identity matrix $W = \mathbf{1}_n$. The loss is then the distance measured by the squared Euclidean distance in the parameter space: $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction: We take W to be the design matrix $W = \Psi^\top$. The associated loss is just $\|\Psi\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{f}^*\|$ This type of loss is usually called *prediction loss*, as it measures the precision with which we could predict future observations from the same data source.

Semiparametric estimation: Let the target of estimation be some subvector $\boldsymbol{\theta}_1^*$ of dimension n_1 of the whole vector $\boldsymbol{\theta}^*$. The profile estimator is defined as $\Pi_1\widetilde{\boldsymbol{\theta}}_m$, where Π_1 is the projector onto the subspace where $\boldsymbol{\theta}_1^*$ lives. The loss we then consider is the squared Euclidean distance of the projections on the subspace:

$$\ell_m \stackrel{\text{def}}{=} \|\Pi_1(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Linear functional estimation The choice of W can be adjusted to estimate any linear functional of the whole parameter vector $\boldsymbol{\theta}^*$. Let us assume that $\boldsymbol{\theta}$ represents the coefficients of f in some orthonormal basis $(\psi_j)_{1\le j\le\infty}$ and for a fixed i with $1 \le i \le n$:

$$I\!E\,(Y_i) = f(x_i)$$

can then be represented as

$$f(x_i) = \sum_{j\ge1} \theta_j \psi_j(x_i).$$

This gives $W = ((\psi_j(x_i))_{j\ge1})^\top$.

Linear inverse problem Assuming that $\boldsymbol{f}^*$ is the evaluation of a function f^* in the design points $(x_i)_{1\le i\le n}$, we can also choose to estimate a derivative of the function f^* in the design points. For the k-th derivative the associated W is

$$W = (\psi_j^{(k)}(x_i))_{i,j})_{1\le i,j\le n}$$

We consider the loss function

$$\ell_m(\widetilde{\boldsymbol{\theta}}_m) = \|W\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{f}^{*(k)}\|,$$

which gives the risk

$$\mathcal{R}_m = I\!E\left(\|W\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{f}^{*(k)}\|\right).$$

A remark is in order here: with this setup we estimate derivative values of the function

$$f_n^*(x) = \sum_{i=1}^{n} \theta_i^* \psi_i(x).$$

If the true function f^* is smooth enough in the sense of the given orthonormal basis, the derivatives of f^* and f_n^* will be close.

Subsequent results for $W\boldsymbol{\theta}$ will be stated in the Euclidean norm, but under typical smoothness assumptions they can be related to L^2-norm bounds for associated features of a true function f^* by using smoothness properties of the estimated objects. In all the above cases, the most important feature of the estimators $W\widetilde{\boldsymbol{\theta}}_m$ is their *linearity*. It simplifies the study of their theoretical properties including the bias-variance decomposition of the risk of $W\widetilde{\boldsymbol{\theta}}_m$. Namely, for the model (3.1.2) with $I\!E\boldsymbol{\varepsilon} = 0$, it holds

$$\begin{aligned} I\!E\widetilde{\boldsymbol{\theta}}_m &= \boldsymbol{\theta}_m^* = \mathcal{S}_m \boldsymbol{f}^*, \\ \mathcal{R}_m &= \|W\big(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*\big)\|^2 + \operatorname{tr}\big(W\mathcal{S}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{S}_m^\top W^\top\big) \\ &= \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2 + \operatorname{tr}\big(W\mathcal{S}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{S}_m^\top W^\top\big). \end{aligned} \tag{3.1.3}$$

Further, it is implicitly assumed that the bias term $\|W(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|^2$ becomes small when m increases. The smallest model $m_{\min}$ usually has a large bias,

while m large ensures a good approximation quality $\boldsymbol{\theta}^*_m \approx \boldsymbol{\theta}^*$ and a small bias at the cost of an increase in complexity measured by the variance term. In the case of projection estimation, the bias term in (3.1.3) describes the accuracy of approximating the response $\boldsymbol{f}^*$ by an m-dimensional linear subspace and this approximation improves as m grows. We will also call $m_{\max}$ the largest model in $\mathcal{M}$. We also write

$$\mathcal{M}^+(m) \stackrel{\text{def}}{=} \{m \in \mathcal{M} \colon m^\circ > m\}\,.$$

for the set of all models larger than the model m.

3.2 The model selection step

Due to the linear structure of the estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ and of the weighting matrix W, one can consider

$$\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y}$$

with $\mathcal{K}_m = W\mathcal{S}_m \colon \mathbb{R}^n \to \mathbb{R}^q$, $m \in \mathcal{M}$, as a family of linear estimators of the q-dimensional target of estimation

$$\boldsymbol{\phi}^* = W\boldsymbol{\theta}^* = W\mathcal{S}\boldsymbol{f}^* = \mathcal{K}\boldsymbol{f}^*$$

for $\mathcal{K} = W\mathcal{S}$.

Now, we explain the variation of Lepski's method we are using for the approach. Suppose that the estimators in $\{\widetilde{\boldsymbol{\phi}}_m\}_{m\in\mathcal{M}}$ can be totally ordered by their complexity (variance). We write this as:

$$\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_m^\top \le \mathcal{K}_{m^\circ} \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_{m^\circ}^\top, \qquad m^\circ > m, m, m^\circ \in \mathcal{M},$$

where we write $\le$ for the semidefinite ordering of matrices. One would like to select the smallest possible model $m \in \mathcal{M}$ which still provides a reasonable fit. The latter means that the bias component

$$\|\boldsymbol{b}_m\|^2 = \|\boldsymbol{\phi}^*_m - \boldsymbol{\phi}^*\|^2 = \|(\mathcal{K}_m - \mathcal{K})\boldsymbol{f}^*\|^2$$

in the risk decomposition (3.1.3) is not significantly larger than the variance

$$\operatorname{tr}\Big(\operatorname{Var}\big(\widetilde{\boldsymbol{\phi}}_m\big)\Big) = \operatorname{tr}\Big(\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_m^\top\Big)\,.$$

If $m^\circ \in \mathcal{M}$ is such a "good" choice, then the ordering assumption yields that a further increase of the index m over m° only increases the complexity (variance) of the method without real gain in the quality of approximation. This latter fact can be interpreted in terms of pairwise comparisons: whatever $m \in \mathcal{M}$ with $m > m^\circ$ we take, there is no significant bias reduction in using a larger model m instead of m°. This leads to a multiple testing procedure: for each pair $m > m^\circ$ from $\mathcal{M}$, we consider a hypothesis of no significant bias between the models m° and m, and let τ_{m,m° be the corresponding test. The model m° is accepted if $\tau_{m,m^\circ} = 0$ for all $m > m^\circ$. Finally, the selected model is the "smallest accepted":

$$\widehat{m} \stackrel{\text{def}}{=} \operatorname{argmin}\bigl\{m^\circ \in \mathcal{M}\colon \tau_{m,m^\circ} = 0,\, \forall m > m^\circ\bigr\}. \tag{3.2.1}$$

Usually the test τ_{m,m° can be written in the form

$$\tau_{m,m^\circ} = \mathbb{1}\bigl\{\mathbb{T}_{m,m^\circ} > \mathfrak{z}_{m,m^\circ}\bigr\} \tag{3.2.2}$$

for some *test statistics* $\mathbb{T}_{m,m^\circ}$ and for *critical values* $\mathfrak{z}_{m,m^\circ}$. Below we define statistics based on the norms of differences $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

$$\begin{aligned} \mathbb{T}_{m,m^\circ} &= \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\|, \\ \mathcal{K}_{m,m^\circ} &\stackrel{\text{def}}{=} \mathcal{K}_m - \mathcal{K}_{m^\circ}. \end{aligned} \tag{3.2.3}$$

The main difference to what we have introduced as the *classical* Lepski's method in Section 2.6 is that we do choose the smallest accepted model and not the smallest model which is accepted and for which all bigger models are accepted too. For the study of a method using the same comparisons in a sequence space setup, we also refer to [Birgé, 2001]. Answering the question of which comparisons to use for the acceptance criterion is just one step in the definition of a Lepski-type method. Next we address the issue of how to choose the critical values $\{\mathfrak{z}_{m,m^\circ}\}_{m,m^\circ\in\mathcal{M}}$. We will propose a general procedure for this choice, which works for all the estimation problems we have introduced above alike. It will be based on the so-called *propagation condition*, similar in spirit to the one from [Spokoiny and Vial, 2009]: if a model m° is "good" in the sense explained above, it has to be accepted with a high probability. This rule can be seen as an analog of a condition

on the family-wise error rate in a multiple testing problem. Rejecting a "good" model is the family-wise error of first kind, and this error has to be controlled.

Oracle choice

To specify precisely what we mean by a "good model", we use below for each pair $m > m^\circ$ from $\mathcal{M}$ the decomposition

$$\mathbb{T}_{m,m^\circ} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\| = \|\mathcal{K}_{m,m^\circ}(\boldsymbol{f}^* + \boldsymbol{\varepsilon})\| = \|\boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|,$$

where with $\mathcal{K}_{m,m^\circ} = \mathcal{K}_m - \mathcal{K}_{m^\circ}$ we write

$$\boldsymbol{b}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{f}^*,$$

for the bias part and

$$\boldsymbol{\xi}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}. \tag{3.2.4}$$

for the stochastic part of the estimator difference. It obviously holds $I\!E\boldsymbol{\xi}_{m,m^\circ} = 0$. We also define

$$\begin{aligned} \boldsymbol{b}_m &\stackrel{\text{def}}{=} \mathcal{K}_m\boldsymbol{f}^*, \\ \boldsymbol{\xi}_m &\stackrel{\text{def}}{=} \mathcal{K}_m\boldsymbol{\varepsilon}. \end{aligned} \tag{3.2.5}$$

We introduce a matrix $\mathbb{V}_{m,m^\circ} \in I\!R^{q\times q}$ as the variance of $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

$$\mathbb{V}_{m,m^\circ} \stackrel{\text{def}}{=} \operatorname{Var}\bigl(\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\bigr) = \operatorname{Var}\bigl(\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\bigr) = \mathcal{K}_{m,m^\circ}\operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m,m^\circ}^\top.$$

If the noise $\boldsymbol{\varepsilon}$ is homogeneous with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{1}_n$, it holds

$$\begin{aligned} \mathbb{V}_{m,m^\circ} &= \sigma^2\,\mathcal{K}_{m,m^\circ}\,\mathcal{K}_{m,m^\circ}^\top \\ \mathbb{V}_m &\stackrel{\text{def}}{=} \sigma^2\,\mathcal{K}_m\,\mathcal{K}_m^\top. \end{aligned}$$

We define the *effective dimensions* of the quadratic forms as:

$$\mathtt{p}_{m,m^\circ} \stackrel{\text{def}}{=} \operatorname{tr}(\mathbb{V}_{m,m^\circ}) = I\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2, \quad \mathtt{p}_m \stackrel{\text{def}}{=} \operatorname{tr}(\mathbb{V}_m) = I\!E\|\boldsymbol{\xi}_m\|^2 \tag{3.2.6}$$

and

$$\lambda_{\mathbb{V}_{m,m^\circ}} \stackrel{\text{def}}{=} \|\mathbb{V}_{m,m^\circ}\|_{\text{op}}, \quad \lambda_{\mathbb{V}_m} \stackrel{\text{def}}{=} \|\mathbb{V}_m\|_{\text{op}},$$

will denote the maximal (in magnitude) eigenvalues of $\mathbb{V}_{m,m^\circ}$ and $\mathbb{V}_m$.

We can then write:

$$I\!\!E\, \mathbb{T}^2_{m,m^\circ} = \|\boldsymbol{b}_{m,m^\circ}\|^2 + I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|\boldsymbol{b}_{m,m^\circ}\|^2 + \mathtt{p}_{m,m^\circ}, \quad (3.2.7)$$

The bias term $\boldsymbol{b}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{f}^*$ is significant, if its squared norm is competitive with the variance term $\mathtt{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ})$. We say that m° is a "good" choice if there is no significant bias $\boldsymbol{b}_{m,m^\circ}$ for any $m > m^\circ$. This condition will be quantified by a bias-variance trade-off similarly to the one in (2.2.1):

$$\|\boldsymbol{b}_{m,m^\circ}\|^2 \le \beta^2\, \mathtt{p}_{m,m^\circ}, \qquad m > m^\circ \quad (3.2.8)$$

for a given parameter β which controls the bias component in the risk due to decomposition (3.2.7). Now we define the *oracle* m^* as the minimal m° with the property (3.2.8):

$$m^* \stackrel{\text{def}}{=} \min\Big\{m^\circ \in \mathcal{M}\colon \max_{m>m^\circ}\big\{\|\boldsymbol{b}_{m,m^\circ}\|^2 - \beta^2\, \mathtt{p}_{m,m^\circ}\big\} \le 0\Big\}. \quad (3.2.9)$$

We have seen in Chapter 2 that this oracle definition is equivalent to the classical one in the sequence space setup for $\beta = 1$, which motivates the definition.

Now we are going to address the central question of how to choose the critical values for the method.

3.3 The calibration step

We are now going to explain the choice of critical values, when the noise distribution is known. Let us assume a Gaussian setup for a moment. Following the ideas of the motivating example in the section before, we say that a model m is *accepted*, if

$$\|W(\widetilde{\boldsymbol{\theta}}_{m^\circ} - \widetilde{\boldsymbol{\theta}}_m)\| \le z^+_{m^\circ,m}(\mathtt{x}) + \beta\sqrt{\mathtt{p}_{m^\circ,m}};\ m^\circ \in \mathcal{M}^+(m),$$

where $z^{+}_{m^\circ,m}(\mathtt{x})$ will be chosen further below to ensure that the oracle model is accepted with high probability. This means that we will choose

$$\mathfrak{z}_{m^\circ,m} = z^{+}_{m^\circ,m}(\mathtt{x}) + \beta\sqrt{\mathtt{p}_{m^\circ,m}} \tag{3.3.1}$$

as our critical values.

In the case of Gaussian errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n)$, knowledge of the noise structure implies that we know the distributions of $\boldsymbol{\xi}_{m^\circ,m}$ for all $m^\circ > m, m^\circ, m \in \mathcal{M}$, which are in fact of the form: $\mathcal{N}(0, \mathbb{V}^2_{m,m^\circ})$. We now introduce, for each pair $m^\circ > m$ in $\mathcal{M}$, a *tail function* $z_{m^\circ,m}(\cdot) : [0,\infty) \to [0,1]$, such that for each $x \geq 0$:

$$I\!P\left(\|\boldsymbol{\xi}_{m^\circ,m}\| > z_{m^\circ,m}(x)\right) = \exp(-x). \tag{3.3.2}$$

Here we assume that the distribution of $\|\boldsymbol{\xi}_{m^\circ,m}\|$ is continuous and that the function $z_{m^\circ,m}(\cdot)$ is thus well-defined. Otherwise we define the value of the tail function for $x \geq 0$ as the smallest value such that the probability is smaller than or equal to $\exp(-x)$. One can see that knowledge of the tail function is just a monotone reparametrization of the quantile function of $\|\boldsymbol{\xi}_{m^\circ,m}\|$. We recall the notation

$$\mathcal{M}^{+}(m) = \{m \in \mathcal{M} \colon m^\circ > m\}.$$

To ensure a propagation condition, we need a uniform in $m^\circ > m$ version of the above probability bound (3.3.2). Sets of concentration for a certain probability level can of course be defined in different ways. We opt for the following parametrization in terms of

$$I\!P\left(\bigcup_{m^\circ \in \mathcal{M}^{+}(m)} \left\{\|\boldsymbol{\xi}_{m^\circ,m}\| \geq z_{m^\circ,m}(\mathtt{x} + q_m)\right\}\right) = \exp(-\mathtt{x}), \tag{3.3.3}$$

where $q_m \geq 0$ is defined by this relation. This means that we construct the uniform concentration sets as enlarged tail functions of the individual $\|\boldsymbol{\xi}_{m^\circ,m}\|$. One simple way to obtain an upper bound on the multiplicity correction q_m is based on the *Bonferroni* bound: As a worst case we can assume independence of all considered variables, which gives $q_m \leq \log\left(|\mathcal{M}^{+}(m)|\right)$. But in many examples of ordered model selection we know, by looking at

the construction of the $\|\boldsymbol{\xi}_{m^\circ,m}\|$, that we will have significant dependency between the different random variables. The Bonferroni correction would be very conservative and would be significantly bigger than q_m. As we can sample from the joint distribution of the $\boldsymbol{\xi}_{m^\circ,m}$'s, we can compute the values for q_m by simulation. We do not know any analytical expression for this type of joint distribution, therefore simulation seems to be the only way obtain the correction terms. Let us denote

$$z^+_{m^\circ,m}(\mathtt{x}) \stackrel{\text{def}}{=} z_{m^\circ,m}(\mathtt{x}+q_m), m^\circ > m.$$

We now summarize the algorithm for calibration: For each $m \in \mathcal{M}$ and a fixed $\mathtt{x} > 0$:

- first approximate the tail functions of $\|\boldsymbol{\xi}_{m^\circ,m}\|$ by simulation and determine $z_{m^\circ,m}(\cdot)$ for all $m^\circ \geq m$: $\forall x \geq 0$:

$$I\!P(\|\boldsymbol{\xi}_{m^\circ,m}\| \geq z_{m^\circ,m}(x)) = \exp(-x).$$

- find a correction term $q_m \geq 0$ to ensure

$$I\!P(\bigcup_{m^\circ \geq m} \{\|\boldsymbol{\xi}_{m^\circ,m}\| \geq z_{m^\circ,m}(\mathtt{x}+q_m)\}) = \exp(-\mathtt{x}).$$

  Choose

$$z^+_{m^\circ,m}(\mathtt{x}) = z_{m^\circ,m}(\mathtt{x}+b_m).$$

- add bias correction to get

$$\mathfrak{z}_{m^\circ,m} = z^+_{m^\circ,m}(\mathtt{x}) + \beta\sqrt{\mathfrak{p}_{m,m^\circ}}.$$

This definition still involves the two numerical tuning constants $\mathtt{x}$ and β. The first value $\mathtt{x}$ controls the nominal rejection probability under the null, a usual choice $\mathtt{x} = 3$ does a good job in most cases. The value β controls the amount of admissible bias in the definition of a good choice; cf. (3.2.8) and (3.2.9). In non-pathological cases, a choice in the range between 0 and 1 normally works well.

In the following section, we will study the optimality of this model selection method.

3.4 An oracle inequality for probabilistic loss

We choose $\widehat{m}$ as explained above by the Smallest-Accepted method. With the $\mathfrak{z}_{m,m^\circ}$'s from (3.3.1), we recall the definition of our model selector: the acceptance rule reads as follows:

$$\widehat{m} \stackrel{\text{def}}{=} \operatorname{argmin}\bigl\{m^\circ \in \mathcal{M}\colon \max_{m \in \mathcal{M}^+(m^\circ)} \bigl\{\mathbb{T}_{m,m^\circ} - \mathfrak{z}_{m,m^\circ}\bigr\} \le 0\bigr\}$$

with $\mathbb{T}_{m,m^\circ}$ defined in (3.2.3). The bound (3.3.3) automatically ensures the desired *propagation property*: any good model m° in the sense (3.2.8) will be accepted with probability at least $1 - \exp(-\mathtt{x})$. In some sense, this property is built-in by construction of the procedure. By definition, the oracle m^* is also a "good" choice, this yields

$$I\!P\bigl(m^* \text{ is rejected}\bigr) \le \exp(-\mathtt{x}). \tag{3.4.1}$$

Therefore, the selector $\widehat{m}$ typically takes its value in $\mathcal{M}^-(m^*)$, where we define

$$\mathcal{M}^-(m^*) = \bigl\{m \in \mathcal{M}\colon m < m^*\bigr\}$$

as the set of all models in $\mathcal{M}$ smaller than m^*. It remains to check the performance of the method in this region. Having a control from above on the location of our model selector, we next define a subset $\mathcal{M}^\circ$ of $\mathcal{M}^-(m^*)$ of possible $\widehat{m}$-values. We will call this subset the *zone of insensitivity*. The definition of m^* implies that there is a significant bias for each $m \in \mathcal{M}^-(m^*)$. If the bias gets large enough, then, again, the probability of selecting m can be bounded from above by a small value. Therefore, the zone of insensitivity is composed of m-values for which the bias is significant but not so large as to dominate completely. This is the subset of $\mathcal{M}$ to which $\widehat{m}$ will belong with high probability. First we define the set where the bias is very large. We recall that $z_{m,m^\circ}(\cdot)$ is the tail function from (3.3.2) for each pair $m > m^\circ \in \mathcal{M}$. We define

$$\mathcal{M}^c = \{m \in \mathcal{M}^-(m^*)\colon \|\boldsymbol{b}_{m^*,m}\| > \mathfrak{z}_{m^*,m} + z_{m^*,m}(\mathtt{x}^c)\}, \tag{3.4.2}$$

where $\mathtt{x}^c \stackrel{\text{def}}{=} \mathtt{x} + \log(|\mathcal{M}^c|)$. The zone of insensitivity is now $\mathcal{M}^\circ \stackrel{\text{def}}{=} \mathcal{M}^-(m^*) \setminus \mathcal{M}^c$. With these definitions, we get the following theorem:

Theorem 3.4.1. *Given* $\mathtt{x}$ *and* β, *let* $\mathfrak{z}_{m,m^\circ}$ *be given by* (3.3.1). *Then the propagation property* (3.4.1) *is satisfied for the SmA selector* $\widehat{m}$.

It also holds

$$I\!P\big(\widehat{m} \in \mathcal{M}^c\big) \leq \exp(-\mathtt{x}).$$

The SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ *satisfies the following bound:*

$$I\!P\Big(\big\|\widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^*}\big\| > \bar{\mathfrak{z}}_{m^*}\Big) \leq 2\exp(-\mathtt{x}), \tag{3.4.3}$$

where $\bar{\mathfrak{z}}_{m^*}$ *is defined as*

$$\bar{\mathfrak{z}}_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^\circ} \mathfrak{z}_{m^*,m}. \tag{3.4.4}$$

This implies the probabilistic oracle bound: on a random set $\Omega(\mathtt{x})$ *with* $I\!P(\Omega(\mathtt{x})) \geq 1 - 2\exp(-\mathtt{x})$ *it holds*

$$\big\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\big\| \leq \big\|\widetilde{\boldsymbol{\phi}}_{m^*} - \boldsymbol{\phi}^*\big\| + \bar{\mathfrak{z}}_{m^*}. \tag{3.4.5}$$

Note that the choice of $\mathtt{x}^c = \mathtt{x} + \log(|\mathcal{M}^c|)$ is based a Bonferroni correction. One could get a smaller set $\mathcal{M}^c$ by choosing $\mathtt{x}^c$ more carefully. But as we only use this value in the theoretical bounds and it is not used in the procedure, a fine tuning for this value is not required. The result (3.4.5) is called the *oracle bound*, because it compares the loss of the data-driven selector $\widehat{m}$ and of the optimal choice m^*. The value $\bar{\mathfrak{z}}_{m^*}$ in (3.4.4) can be viewed as the "payment for adaptation". An interesting feature of the presented result is that not only the oracle quality but also the payment for adaptation depend upon the unknown response $\boldsymbol{f}^*$ and the corresponding oracle choice m^*. In the worst case of the model with a flat enough risk profile $\mathcal{R}_m$, the set $\mathcal{M}^\circ$ can coincide with the whole range $\mathcal{M}^-(m^*)$. Even in this case the bounds (3.4.3) and (3.4.5) are meaningful. However, the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ in this case can be larger than the oracle risk. On the other hand, if the risk function $\mathcal{R}_m$ grows fast enough as m decreases below m^*, then the set $\mathcal{M}^\circ$ is small and the value $\bar{\mathfrak{z}}_{m^*}$ is much smaller than the oracle risk $\mathcal{R}_{m^*}$.

Analysis of the payment for adaptation

In the following section, we are going to study the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ more closely for the special case of an independent Gaussian error vector $\boldsymbol{\varepsilon}$. In this case, one can heavily rely on the nice properties of Gaussian random variables under linear transformations. However, the results should be extendable to the case of non-Gaussian errors $\boldsymbol{\varepsilon}$ under subgaussian moment conditions. We write $\mathbb{V}_{m,m^\circ} = \mathrm{Var}(\boldsymbol{\xi}_{m,m^\circ})$ for the covariance matrices of the $\boldsymbol{\xi}_{m,m^\circ}$. We recall $\mathtt{p}_{m,m^\circ} = \mathrm{tr}(\mathbb{V}_{m,m^\circ})$ and $\lambda_{m,m^\circ} = \|\mathbb{V}_{m,m^\circ}\|_{\mathrm{op}}$.

Theorem 3.4.2. *Assume the conditions of Theorem 3.4.1 and let* $\mathtt{p}_{m^*,m} \leq \mathtt{p}_{m^*,m_{min}} \leq \mathtt{p}_{m^*}$ *and* $\lambda_{m^*,m} \leq \lambda_{m^*,m_{min}} \leq \lambda_{m^*}$ *for all* $m_{min} \leq m < m^*$. *If the errors* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n)$, $\sigma > 0$, *are Gaussian, then the critical values* $\mathfrak{z}_{m,m^\circ}$ *given by* (3.3.1) *satisfy*

$$\mathfrak{z}_{m,m^\circ} \leq (1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\{\mathtt{x} + \log(|\mathcal{M}|)\}},$$

while for the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ *the following bound holds: for the smallest model* $m_{min} \in \mathcal{M}^\circ$

$$\begin{aligned}\bar{\mathfrak{z}}_{m^*} &\leq (1+\beta)\sqrt{\mathtt{p}_{m^*,m_{min}}} + \sqrt{2\lambda_{m^*,m_{min}}\{\mathtt{x} + \log(|\mathcal{M}^-(m^*)|)\}} \\ &\leq (1+\beta)\sqrt{\mathtt{p}_{m^*}} + \sqrt{2\lambda_{m^*}\{\mathtt{x} + \log(|\mathcal{M}|)\}}\,.\end{aligned}$$

Applications of this result to the case of the estimation of a whole function vector and the case of linear functional estimation will be discussed in Sections 3.6.1 and 3.6.2 below.

3.5 An oracle inequality for a polynomial loss function

In the setup before, we have calibrated our acceptance bounds by using a probabilistic loss. We are now trying to construct acceptance bounds which are also sensible to the magnitude of the loss. In the following, we will the consider quadratic risk:

$$\mathcal{R}(\theta) \stackrel{\mathrm{def}}{=} I\!E\left(\|W(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2\right).$$

The aim now is to bound the risk of the Smallest-Accepted method by the risk of an oracle estimator $W\boldsymbol{\theta}_{m^*}$. We rewrite

$$W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = \boldsymbol{\xi}_m + \boldsymbol{b}_m.$$

We look at the risk

$$\mathcal{R}_m \stackrel{\text{def}}{=} I\!E(\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2) = I\!E(\|\boldsymbol{\xi}_m\|^2) + \|\boldsymbol{b}_m\|^2 = \mathtt{p}_m + \|\boldsymbol{b}_m\|^2.$$

For the risk bound we will modify the definition of the oracle slightly

$$m^* \stackrel{\text{def}}{=} \min\left\{ m \in \mathcal{M}\colon \max_{m',m^\circ \in \mathcal{M}^+(m):m'>m^\circ} \left\{ \|\boldsymbol{b}_{m',m^\circ}\|^2 - \beta^2 \mathtt{p}_{m',m^\circ} \right\} \le 0 \right\} \tag{3.5.1}$$

We also assume that the bias satisfies:

$$\|\boldsymbol{b}_m\| \le \|\boldsymbol{b}_{m^*}\|, m > m^*. \tag{3.5.2}$$

Otherwise one defines $\|\boldsymbol{b}_{m^*}\| \stackrel{\text{def}}{=} \max_{m\in\mathcal{M}^+(m^*)} \|\boldsymbol{b}\|$. We also assume that our set of models $\mathcal{M}$ is finite and assume, as before, that we have a total order on the models. We then write $m-1$ for the largest model which is smaller than m.

We define

$$\mathcal{R}^+(\mathtt{x}) \stackrel{\text{def}}{=} I\!E\left[\left(\frac{\|\boldsymbol{\xi}_m\|^2}{\mathtt{p}_m} \vee 1 \right) \mathbb{1}\left(\max_{m^\circ \in \mathcal{M}^+(m-1)} \left(\|\boldsymbol{\xi}_{m^\circ,m-1}\| - z_{m^\circ,m-1}(\mathtt{x}) \right) > 0 \right) \right]$$

and we abbreviate

$$A_m \stackrel{\text{def}}{=} \{ \max_{m^\circ \in \mathcal{M}^+(m)} \left(\|\boldsymbol{\xi}_{m^\circ,m}\| - z_{m^\circ,m-1}(\mathtt{x}) \right) > 0 \}.$$

We note that by replacing $\left(\frac{\|\boldsymbol{\xi}_m\|}{\sqrt{\mathtt{p}_m}}\right)^2$ by $\left(\frac{\|\boldsymbol{\xi}_m\|}{\sqrt{\mathtt{p}_m}}\right)^q$, we can treat any other polynomial loss with exponent $q > 0$ and as a specific case for $q = 0$ we recover the condition for probabilistic loss. For all $m \in \mathcal{M}\backslash m_{\min}$, we define $\mathtt{x}_{m-1}$ by the relation

$$\mathcal{R}_m^+(\mathtt{x}_{m-1}) = \alpha_m. \tag{3.5.3}$$

This gives

$$\begin{aligned} I\!E\left(\|\boldsymbol{\xi}_m\|^2 \mathbb{1}(A_{m-1}(\mathtt{x}_{m-1})\right) &\leq \alpha_m \mathtt{p}_m, \\ I\!P\left(A_{m-1}(\mathtt{x}_{m-1})\right) &\leq \alpha_m. \end{aligned} \tag{3.5.4}$$

We then define the critical values for the procedure as

$$\mathfrak{z}_{m^\circ,m} \stackrel{\text{def}}{=} z_{m^\circ,m}(\mathtt{x}_m) + \beta \mathtt{p}_{m^\circ,m}^{1/2}. \tag{3.5.5}$$

The general model selection step of the Smallest-Accepted method stays the same and all that changes are the critical values.

Theorem 3.5.1. *Let the SmA procedure* (3.2.1) *be applied with the critical values* $\mathfrak{z}_{m,m^\circ}$ *from* (3.5.5)*, where the values* $\mathtt{x}_m$ *are defined by* (3.5.3) *with the coefficients* α_m *satisfying*

$$\sum_{m \in \mathcal{M}^+(m^*)} \alpha_m \mathtt{p}_m \leq \overline{\alpha}_{m^*} \mathtt{p}_{m^*} \tag{3.5.6}$$

for some $\overline{\alpha}_{m^*}$ *. If the errors* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n)$ *,* $\sigma > 0$*, are Gaussian, then*

$$I\!E\left(\left\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right\|^2\right) \leq 2\overline{\alpha}_{m^*} \mathcal{R}_{m^*} + \left(\mathcal{R}_{m^*}^{1/2} + \overline{\mathfrak{z}}_{m^*}\right)^2, \tag{3.5.7}$$

where

$$\overline{\mathfrak{z}}_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^-(m^*)} \mathfrak{z}_{m^*,m}\,.$$

Now we discuss the choice of the constants α_m entering in the definition (3.5.3). Suppose that the $\mathtt{p}_m$'s satisfy

$$\sum_{m \in \mathcal{M}^+(m^*)} (\mathtt{p}_{m^*}/\mathtt{p}_m)^a \leq C \tag{3.5.8}$$

for some $a > 0$ and a fixed constant C. Then one can take

$$\alpha_m = \mathtt{p}_m^{-1-a}$$

yielding

$$\sum_{m \in \mathcal{M}^+(m^*)} \alpha_m \mathtt{p}_m \leq \sum_{m \in \mathcal{M}^+(m^*)} \mathtt{p}_m^{-a} \leq C \mathtt{p}_{m^*}^{-a} = C \overline{\alpha}_{m^*} \mathtt{p}_{m^*}$$

with $\overline{\alpha}_{m^*} = C\mathtt{p}_{m^*}^{-1-a}$. In the next proposition we study the situation when the effective dimension $\mathtt{p}_m$ grows exponentially in m. Then (3.5.8) is satisfied for any $a > 0$ with $\mathtt{C} = \mathtt{C}(a)$.

The next step is an upper bound on the values $\mathtt{x}_m$, $z_{m,m^\circ}(\mathtt{x}_m)$, and $\mathfrak{z}_{m,m^\circ}$, as well as on the payment for adaptation $\overline{\mathfrak{z}}_{m^*}$.

Proposition 3.5.2. *Suppose* (3.5.8) *for* $a > 0$. *If* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n)$, $\sigma > 0$, *then the choice*

$$\alpha_m = \sqrt{3}(\mathtt{p}_m/\mathtt{p}_{m_{min}})^{-1-a}, \quad \mathtt{x}_{m-1} = 2(1+a)\log(\mathtt{p}_m/\mathtt{p}_{m_{min}}), \tag{3.5.9}$$

ensures conditions (3.5.6), (3.5.3), *and therefore, the oracle bound* (3.5.7) *holds with* $\overline{\alpha}_{m^*} = \sqrt{3}\mathtt{C}(\mathtt{p}_{m_{min}}/\mathtt{p}_{m^*})^{1+a}$. *Furthermore,*

$$\overline{\mathfrak{z}}_{m^*} \le \beta\sqrt{\mathtt{p}_{m^*}} + \sqrt{2\lambda_{m^*}\,(2(1+a)\log(\mathtt{p}_{m^*}/\mathtt{p}_{m_{min}}) + \log(|\mathcal{M}|))}. \tag{3.5.10}$$

In the next section, we are going to apply our method to the case of prediction of the whole function and the estimation of a linear functional.

These two situations are in some sense extreme cases of the relation between $\mathtt{p}_{m^*}$ and λ_{m^*}.

3.6 Examples

3.6.1 Prediction of the whole function

This section discusses the case of projection estimation in the linear model

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$$

with homoscedastic noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma \mathbf{1}_n)$. All the conclusions can be extended to a more general diagonal covariance matrix Σ whose coefficients are lower and upper bounded independent of n. We will state our results for the probabilistic loss. The case of a polynomial loss can be treated analogously. We write $\Psi_m^\top$ for the design matrix associated only with the features selected by the model m. We also write $\widetilde{\boldsymbol{\theta}}_m$ for the least squares estimator

associated with

$$\boldsymbol{Y} = \Psi_m^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

If we choose $W = \Psi^\top$, we get $\mathcal{S}_m = \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m$ $\mathcal{K}_m \boldsymbol{Y} = \Pi_m \boldsymbol{Y}$ where $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m$ is a projector on the subspace spanned by the features in m. In the homoscedastic setting we have

$$\mathtt{p}_m = \operatorname{tr}\bigl(\operatorname{Var}\bigl(\Pi_m \boldsymbol{Y}\bigr)\bigr) = \sigma^2 \operatorname{tr}\bigl(\Pi_m\bigr) = \sigma^2 m.$$

Moreover, for each pair $m > m^\circ$, it holds

$$\Psi^\top \bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) = \bigl(\Pi_m - \Pi_{m^\circ}\bigr) \boldsymbol{Y} = \Pi_{m,m^\circ} \boldsymbol{Y},$$

where Π_{m,m° projects on the subspace of features which belong to m but not to m°. In this setting, Theorem 3.4.2 gives the following corollary.

Corollary 3.6.1. *In the setting above, we have* $\mathtt{p}_{m,m^\circ} = \sigma^2 (m - m^\circ)$, $\lambda_{m,m^\circ} = \sigma^2$, *and*

$$\begin{aligned} \mathfrak{z}_{m,m^\circ} &\le \sigma(1+\beta)\sqrt{m - m^\circ} + \sigma\sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}, \\ \bar{\mathfrak{z}}_{m^*} &\le \sigma(1+\beta)\sqrt{m^*} + \sigma\sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}. \end{aligned}$$

The first term in the bound for $\bar{\mathfrak{z}}_{m^*}$ is of order $\sqrt{m^*}$ and it should be the dominating term, when $\mathtt{p}_{m^*}$ is significantly larger than $\log(|\mathcal{M}|)$. Usually, one can choose the set $\mathcal{M}$ to be of cardinality of order $\log(n)$; cf. [Lepski, 1991, Lepski et al., 1997]. In this situation $\log(|\mathcal{M}|)$ will be of order $\log(\log(n))$ and $\sigma\sqrt{m^*}$ will be the dominating term for $m^* \gg \log(\log(n))$. For the oracle risk $\mathcal{R}_{m^*}$, it holds $\mathcal{R}_{m^*} = \mathtt{p}_{m^*} + \|\boldsymbol{b}_{m^*}\|^2 \ge \sigma^2 m^*$. Therefore, the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ is of the same order as the square root of the oracle risk, and the result of Proposition 3.4.2 has an interesting corollary: rate-adaptive estimation is possible if the oracle dimension m^* is significantly larger than $\log(\log(n))$.

It is possible to get a sharper bound on the payment for adaptation if the zone of insensitivity is small. If the bias grows rapidly when m decreases from m^* to $m_{\min}$, more precisely, if $\|\boldsymbol{b}_{m^*,m}\|^2 \ge \mathtt{C}\sigma^2 \bigl(m^* - m +$

$2\mathtt{x} + 2\log(|\mathcal{M}|)\big)$ for a fixed constant $\mathtt{C} > 0$ and for all $m \le m^\circ$ such that $m^\circ < m^*$, then

$$\bar{\mathfrak{z}}_{m^*} \le \sigma(1+\beta)\sqrt{m^* - m^\circ} + \sigma\sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}.$$

This means that if the ratio $(m^* - m^\circ)/m^*$ is small, the payment for adaptation is smaller in order than the oracle risk, and the procedure is sharp adaptive in probabilistic loss. One can use a condition like the self-similarity condition introduced in [Giné and Nickl, 2010] to assure that the bias grows fast enough.

3.6.2 Linear functional estimation

Now we are going to treat the case of the estimation of a linear functional $W \in I\!R^{1\times n}$. Other than changing W, we take the same setup as in the previous section. We again write

$$\widetilde{\phi}_m = W\widetilde{\boldsymbol{\theta}}_m = \mathcal{K}_m \boldsymbol{Y}, \quad m \in \mathcal{M}. \tag{3.6.1}$$

We assume for simplicity that m also denotes the number of features. The ordering condition means that these estimators are ordered by their variance:

$$\mathrm{v}_m^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\big(\mathcal{K}_m \boldsymbol{Y}\big) = \mathcal{K}_m \, \mathrm{Var}(\boldsymbol{\varepsilon}) \, \mathcal{K}_m^\top$$

which grows with m and that the bias decreases for growing m. Further, each stochastic component $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}$ is one-dimensional, and it holds

$$\lambda_{m,m^\circ} = \mathtt{p}_{m,m^\circ} = \mathrm{v}_{m,m^\circ}^2 = \mathcal{K}_{m,m^\circ} \, \mathrm{Var}(\boldsymbol{\varepsilon}) \, \mathcal{K}_{m,m^\circ}^\top.$$

Note that in the case of Gaussian errors, ξ_{m,m° is also Gaussian: $\xi_{m,m^\circ} \sim \mathcal{N}(0, \mathrm{v}_{m,m^\circ}^2)$. The tail function $z_{m,m^\circ}(\mathtt{x})$ of ξ_{m,m° can be upper-bounded by $\mathrm{v}_{m,m^\circ}\sqrt{2\mathtt{x}}$. In the case of probabilistic loss, a Bonferroni correction and a bias adjustment lead to the following upper bound for the critical values $\mathfrak{z}_{m,m^\circ}$:

$$\mathfrak{z}_{m,m^\circ} \le \mathrm{v}_{m,m^\circ}\Big(\beta + \sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}\Big). \tag{3.6.2}$$

This implies

$$\bar{\mathfrak{z}}_{m^*} \le \mathrm{v}_{m^*}\Big(\beta + \sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}\Big).$$

We summarize this in the following corollary.

Corollary 3.6.2. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{1}_n), \sigma > 0$. *Consider the target of estimation* $\phi^* = \mathcal{K}\boldsymbol{f}^*$ *and the family of one-dimensional estimators* $\{\widetilde{\phi}_m\}_{m\in\mathcal{M}}$ *defined in* (3.6.1). *Then the critical values* $\mathfrak{z}_{m,m^\circ}$ *from* (3.3.1) *satisfy* (3.6.2) *and the oracle inequality* (3.4.5) *holds with the payment for adaptation*

$$\bar{\mathfrak{z}}_{m^*} \le \mathrm{v}_{m^*}\Big(\beta + \sqrt{2\mathtt{x} + 2\log(|\mathcal{M}|)}\Big).$$

One can conclude that for the problem of functional estimation with probabilistic loss, the squared payment for adaptation $\bar{\mathfrak{z}}^2_{m^*}$ is by a factor $\log(|\mathcal{M}|)$ larger than the oracle variance $\mathrm{v}^2_{m^*}$. If $|\mathcal{M}|$ itself is logarithmic in the sample size n, we end up with the extra $\log(\log n))$-factor in the accuracy of adaptive estimation.

In the case of a polynomial loss function, similar arguments yield due to (3.5.5) and (3.5.9)

$$\begin{aligned}\mathfrak{z}_{m,m^\circ} &\le \mathrm{v}_{m,m^\circ}\big(\beta + \sqrt{2\mathtt{x}_{m^\circ} + 2\log(|\mathcal{M}|)}\big)\\ &\le \mathrm{v}_{m,m^\circ}\big(\beta + \sqrt{2(1+a)\log(\mathtt{p}_{m^\circ}/\mathtt{p}_{m_{\min}}) + 2\log(|\mathcal{M}|)}\big)\end{aligned}$$

The bound (3.6.2) yields

$$\bar{\mathfrak{z}}_{m^*} \le \mathrm{v}_{m^*}\Big(\beta + \sqrt{2(1+a)\log(\mathrm{v}^2_{m^*}/\mathrm{v}^2_{m_{\min}}) + 2\log(|\mathcal{M}|)}\Big).$$

It appears that polynomial loss yields a larger payment for adaptation: $\bar{\mathfrak{z}}^2_{m^*} = \mathcal{O}(\mathrm{v}^2_{m^*}\log(\mathrm{v}^2_{m^*}/\mathrm{v}^2_{m_{\min}}))$. This conclusion is consistent with the results by [Lepski, 1992] and [Cai and Low, 2003, Cai and Low, 2005], which show that the log-price for adaptation cannot be avoided if a polynomial loss is considered.

In the next section we give the proofs for the results in this chapter.

3.7 Proofs

3.7.1 Proof of Theorem 3.4.1

By the propagation property (3.4.1) we can be sure that the oracle model m^* will be accepted with high probability. That means that $\widehat{m}$ will not be larger than m^*, that is, $\widehat{m} \le m^*$ with a probability at least $1 - \exp(-\mathtt{x})$. Below we consider only this event. Let $m \in \mathcal{M}^-(m^*)$. If m is accepted, we have by the acceptance criterion (3.2.2) that $\mathbb{T}_{m^*,m} \le \mathfrak{z}_{m^*,m}$. The representation $\mathbb{T}_{m^*,m} = \|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\|$ implies

$$I\!P\bigl(\mathbb{T}_{m^*,m} < \mathfrak{z}_{m^*,m}\bigr) \le I\!P\bigl(\|\boldsymbol{\xi}_{m^*,m}\| > \|\boldsymbol{b}_{m^*,m}\| - \mathfrak{z}_{m^*,m}\bigr).$$

Under (3.4.2) this yields

$$\begin{aligned} I\!P\bigl(m \text{ is accepted}\bigr) &\le I\!P\bigl(\bigl\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\bigr\| \le \mathfrak{z}_{m^*,m}\bigr) \\ &\le I\!P\bigl(\bigl\|\boldsymbol{\xi}_{m^*,m}\bigr\| \ge z_{m^*,m}(\mathtt{x}^c)\bigr) \le \mathrm{e}^{-\mathtt{x}^c}. \end{aligned} \tag{3.7.1}$$

If the lower bound on the bias is satisfied for all $m \in \mathcal{M}^c$, then (3.7.1) helps to bound the probability of the event $\{\widehat{m} \in \mathcal{M}^c\}$:

$$I\!P\bigl(\widehat{m} \in \mathcal{M}^c\bigr) \le \sum_{m \in \mathcal{M}^c} I\!P\bigl(\bigl\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\bigr\| < \mathfrak{z}_{m^*,m}\bigr) \le \sum_{m \in \mathcal{M}^c} \mathrm{e}^{-\mathtt{x}^c} \le \exp(-\mathtt{x}).$$

Therefore, the probability that the SmA-selector selects a model $m > m^*$ or $m \in \mathcal{M}^c$ is bounded by:

$$I\!P\Bigl(\widehat{m} \in \mathcal{M}^+(m^*) \cup \mathcal{M}^c\Bigr) \le 2\exp(-\mathtt{x}).$$

It remains to study the case when $\widehat{m} = m \in \mathcal{M}^\circ = \mathcal{M}^-(m^*) \setminus \mathcal{M}^c$. We can use that $\widehat{m}$ is accepted which implies by definition

$$\mathbb{T}_{m^*,m} = \bigl\|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*}\bigr\| \le \mathfrak{z}_{m^*,m}.$$

This yields (3.4.3). The bound (3.4.5) follows by the triangle inequality.

3.7.2 Proof of Proposition 3.4.2

Below we use the deviation bound (5.2.1) for a Gaussian quadratic form from Theorem 5.2.1. Note that similar results are available for non-Gaussian

quadratic forms under exponential moment conditions; see e.g. [Spokoiny and Zhilova, 2013], [Hsu et al., 2012], [Hanson and Wright, 1971], [Rudelson and Vershynin, 2013]. The result (5.2.1) combined with the Bonferroni correction $q_{m^\circ} = \log(|\mathcal{M}^+(m^\circ)|) \leq \log(|\mathcal{M}|)$ yields the following upper bound for the critical values $\mathfrak{z}_{m,m^\circ}$:

$$\begin{aligned}
\mathfrak{z}_{m,m^\circ} &\leq z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) + \beta \mathtt{p}_{m,m^\circ}^{1/2} \\
&\leq (1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\{\mathtt{x} + \log(|\mathcal{M}^+(m^\circ)|)\}} \\
&\leq (1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\{\mathtt{x} + \log(|\mathcal{M}|)\}}. \qquad (3.7.2)
\end{aligned}$$

For the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$, the result (3.7.2) and the monotonicity condition $\mathtt{p}_{m^*,m} \leq \mathtt{p}_{m^*,m_{\min}} \leq \mathtt{p}_{m^*}$ and $\lambda_{m^*,m} \leq \lambda_{m^*,m_{\min}} \leq \lambda_{m^*}$ imply the following upper bound:

$$\begin{aligned}
\bar{\mathfrak{z}}_{m^*} &\leq (1+\beta)\sqrt{\mathtt{p}_{m^*,m_{\min}}} + \sqrt{2\lambda_{m^*,m_{\min}}\{\mathtt{x} + \log(|\mathcal{M}^-(m^*)|)\}} \\
&\leq (1+\beta)\sqrt{\mathtt{p}_{m^*}} + \sqrt{2\lambda_{m^*}\{\mathtt{x} + \log(|\mathcal{M}|)\}}
\end{aligned}$$

which yields the claim.

3.7.3 Proof of Theorem 3.5.1

The result will be proved in two steps. First we bound the risk on the set $\widehat{m} > m^*$:

$$\mathbb{E}\left(\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|^2 \mathbb{1}(\widehat{m} > m^*)\right) \leq 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*}\,. \qquad (3.7.3)$$

Then we consider the region $\widehat{m} < m^*$ and prove an oracle inequality

$$\|\widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^*}\| \mathbb{1}(\widehat{m} < m^*) \leq \bar{\mathfrak{z}}_{m^*} \qquad (3.7.4)$$

and the oracle bound (3.5.7). We start by proving (3.7.3). Let us fix $m \in \mathcal{M}^+(m^*)$ and $m' \geq m$. The definition (3.5.1) of the oracle m^* and the formula (3.5.5) for the critical value $\mathfrak{z}_{m',m-1}$ implies for the test statistic $\mathbb{T}_{m',m-1} = \|\boldsymbol{\xi}_{m',m-1} + \boldsymbol{b}_{m',m-1}\|$

$$\left\{\mathbb{T}_{m',m-1} > \mathfrak{z}_{m',m-1}\right\} \subseteq \left\{\|\boldsymbol{\xi}_{m',m-1}\| > z_{m',m-1}(\mathtt{x}_{m-1})\right\}.$$

Now we can bound the risk of $\widehat{\boldsymbol{\phi}}$ on the set $\widehat{m} > m^*$. We use that for $\widehat{m} = m > m^*$ we have, in view of (3.5.2),

$$\begin{aligned}\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|^2 &= \|\widetilde{\boldsymbol{\phi}}_m - \boldsymbol{\phi}^*\|^2 = \|\boldsymbol{\xi}_m + \boldsymbol{b}_m\|^2 \\ &\le 2\|\boldsymbol{\xi}_m\|^2 + 2\|\boldsymbol{b}_m\|^2 \le 2\|\boldsymbol{\xi}_m\|^2 + 2\|\boldsymbol{b}_{m^*}\|^2\end{aligned}$$

and it holds by (3.5.4) and monotonicity $\mathtt{p}_m > \mathtt{p}_{m^*}$

$$\begin{aligned}&\mathbb{E}\left(\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|^2 \mathbb{1}\big(\widehat{m} > m^*\big)\right) \\ &\quad\le 2 \sum_{m \in \mathcal{M}^+(m^*)} \mathbb{E}\left(\big(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2\big)\, \mathbb{1}(\widehat{m} = m)\big)\right) \\ &\quad\le 2 \sum_{m \in \mathcal{M}^+(m^*)} \mathbb{E}\left(\big(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2\big)\, \mathbb{1}(m-1 \text{ is rejected})\big)\right) \\ &\quad= 2 \sum_{m \in \mathcal{M}^+(m^*)} \mathbb{E}\Big[(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2) \cdot \\ &\qquad\qquad \mathbb{1}\Big(\max_{m' \in \mathcal{M}^+(m)} \Big\{\|\boldsymbol{\xi}_{m',m-1}\| - z_{m',m-1}(\mathtt{x}_m)\Big\} > 0\Big)\Big] \\ &\quad\le 2 \sum_{m \in \mathcal{M}^+(m^*)} \alpha_m(\mathtt{p}_m + \|\boldsymbol{b}_{m^*}\|^2) \le 2\overline{\alpha}_{m^*}\big(\mathtt{p}_{m^*} + \|\boldsymbol{b}_{m^*}\|^2\big) = 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*}.\end{aligned}$$

Here we have used that (3.5.6) and $\mathtt{p}_m \ge \mathtt{p}_{m^*}$ imply $\sum_{m \in \mathcal{M}^+(m^*)} \alpha_m \le \overline{\alpha}_{m^*}$. This completes the proof of (3.7.3).

In the situation when $\widehat{m} = m < m^*$, we can use the propagation property: as m is accepted, it holds

$$\|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*}\|\, \mathbb{1}(\widehat{m} = m) \le \mathfrak{z}_{m^*,m}\,,$$

which implies (3.7.4) by definition of $\bar{\mathfrak{z}}_{m^*}$. This yields

$$\begin{aligned}\mathbb{E}\left(\big\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\big\|^2\right) &\le 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \mathbb{E}\left(\big\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\big\|^2 \mathbb{1}(\widehat{m} < m^*)\right) \\ &\le 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \mathbb{E}\left(\big(\|\widetilde{\boldsymbol{\phi}}_{m^*} - \boldsymbol{\phi}^*\| + \bar{\mathfrak{z}}_{m^*}\big)^2\right) \\ &\le 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \big(\mathcal{R}_{m^*}^{1/2} + \bar{\mathfrak{z}}_{m^*}\big)^2\end{aligned}$$

as required.

3.7.4 Proof of Proposition 3.5.2

Observe first that the choice $\alpha_m = (\mathtt{p}_m/\mathtt{p}_{m_{\min}})^{-1-a}$ yields

$$\sum_{m \in \mathcal{M}^+(m^*)} \alpha_m \mathtt{p}_m \le \mathtt{p}_{m_{\min}}^{1+a} \sum_{m \in \mathcal{M}^+(m^*)} \mathtt{p}_m^{-a} \le \mathtt{C} \mathtt{p}_{m^*}^{-a} \mathtt{p}_{m_{\min}}^{1+a} = \mathtt{C} \overline{\alpha}_{m^*} \mathtt{p}_{m^*}$$

with $\overline{\alpha}_{m^*} = \mathtt{C}(\mathtt{p}_{m_{\min}}/\mathtt{p}_{m^*})^{1+a}$.

For any random vector $\boldsymbol{\xi}$ with $\mathrm{Var}(\boldsymbol{\xi}) = B$ and $\mathtt{p} = \mathrm{tr}(B)$ and any random event A,

$$\mathbb{E}\Big[\mathtt{p}^{-1}\|\boldsymbol{\xi}\|^2 \mathbb{1}(A)\Big] \le \big\{1 + \mathtt{p}^{-2}\,\mathrm{Var}(\|\boldsymbol{\xi}\|^2)\big\}^{1/2} \mathbb{P}^{1/2}(A). \tag{3.7.5}$$

Indeed, the Cauchy-Schwartz inequality implies

$$\begin{aligned} \mathbb{E}\left(\mathtt{p}^{-1}\|\boldsymbol{\xi}\|^2 \mathbb{1}(A)\right) &\le \mathbb{E}^{1/2}\left(\mathtt{p}^{-1}\|\boldsymbol{\xi}\|^2\right)^2 \mathbb{P}^{1/2}(A) \\ &= \big(1 + \mathtt{p}^{-2}\,\mathrm{Var}(\|\boldsymbol{\xi}\|^2)\big)^{1/2} \mathbb{P}^{1/2}(A) \end{aligned}$$

Moreover, in the Gaussian case $\boldsymbol{\xi} \sim \mathcal{N}(0, B)$ with $\|B\|_{\mathrm{op}} \le 1$, it holds $\mathrm{Var}(\|\boldsymbol{\xi}\|^2) \le 2\mathtt{p}$. If $\mathtt{p}$ is large, then $\mathrm{Var}(\|\boldsymbol{\xi}\|^2)/\mathtt{p}^2$ is small. In general $\mathrm{Var}(\|\boldsymbol{\xi}\|^2)/\mathtt{p}^2 \le 2$.

Result (3.7.5) and the choice $\alpha_m = \sqrt{3}\mathtt{p}_m^{-1-a}$ allow us to specify an upper bound for $\mathtt{x}_m$. Namely, the choice $\mathtt{x}_m = \mathtt{C}\log(\mathtt{p}_m)$ ensures the propagation condition (3.5.3). To see this, fix $m \in \mathcal{M}$ and $m' \ge m$. Let

$$A'_m(\mathtt{x}) \stackrel{\mathrm{def}}{=} \mathbb{1}\Big(\max_{m' \in \mathcal{M}^+(m)} \big\{\|\boldsymbol{\xi}_{m',m}\| - \sqrt{\mathtt{p}_{m',m}} - \sqrt{2\lambda_{m',m}\,\{\mathtt{x} + \log(|\mathcal{M}|)\}}\big\} > 0\Big)$$

The arguments after Lemma 5.2.1 with $\mathtt{x}_{m-1} = 2(1+a)\log(\mathtt{p}_m)$ and (3.7.5) imply

$$\mathbb{E}\Big[\mathtt{p}_m^{-1}\|\boldsymbol{\xi}_m\|^2 \mathbb{1}\{A'_{m-1}(\mathtt{x}_{m-1})\}\Big] \le \sqrt{3}\mathrm{e}^{-(1+a)\log(\mathtt{p}_m)} = \sqrt{3}\mathtt{p}_m^{-1-a}$$

and by (3.5.5)

$$\mathfrak{z}_{m,m^\circ} \le \sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\,\{(1+a)\log(\mathtt{p}_{m^\circ+1}) + \log(|\mathcal{M}|)\}}\,.$$

This implies the upper bound (3.5.10) on the payment for adaptation $\overline{\mathfrak{z}}_{m^*}$.

Chapter 4

Bootstrap-based Smallest-Accepted method

In the first section, we are going to give the setup and some notations for our bootstrap-based method. Then we are presenting the modified calibration step with bootstrapped quantities in the following section. Finally in the last two sections, we are first going to study theoretical properties of the method and show that the results of the previous chapter can essentially be carried over to the new setup, then we illustrate the performance of the method by means of numerical simulations.

4.1 Bootstrap setup

The procedure we are proposing will be related to the concept of the *wild* bootstrap, [Wu, 1986], [Beran, 1986]. The wild bootstrap in the framework of a heteroscedastic regression problem with normal errors proposes to use resampled randomly weighted residuals of an estimator as a replacement for an unknown heteroscedastic error distribution. We will quickly explain the idea in the case of normal errors and normal weights. For different possible weights see for example [Mammen, 1993]. Now to explain the idea, we assume given observations:

$$\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{f}^* \in I\!R^n$ the signal, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ with an unknown diagonal covariance matrix $\Sigma \in I\!R^{n\times n}$. Now given an estimator $\widetilde{\boldsymbol{Y}}$ of $\boldsymbol{f}^*$, we define the residuals:

$$\breve{\boldsymbol{Y}} \stackrel{\text{def}}{=} \boldsymbol{Y} - \widetilde{\boldsymbol{Y}}.$$

We take $\widetilde{\boldsymbol{Y}}$ as a replacement for $\boldsymbol{f}^*$ and we also want to find a bootstrap-proxy of the error distribution $\mathcal{N}(0, \Sigma)$. The wild bootstrap proposes to use conditional on $\breve{\boldsymbol{Y}}$:

$$\mathcal{N}(0, \operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}})), \tag{4.1.1}$$

where we write $\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}}$ for the coordinate-wise product of the vector $\breve{\boldsymbol{Y}}$ with itself and $\operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}})$ denotes the diagonal matrix with entries from $\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}}$. The use of (4.1.1) amounts to multiplying the residuals by normal weights to get the bootstrap-approximation for the error distribution. Of course, in itself $\operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}})$ is usually a very bad estimator for the covariance matrix. But it can turn out, as we will see for our problem of the calibration of critical values, that this replacement can still be useful.

Coming back to our model selection problem, we will use almost the same setup as in the known variance case. Assume that we observe:

$$\boldsymbol{Y} = \boldsymbol{f}^* + \Sigma^{1/2}\boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n),$$

with $\Sigma \in I\!R^{n\times n}$ a positive diagonal matrix and $\boldsymbol{f}^* \in I\!R^n$. We want to choose between different models $m \in \mathcal{M}$, following a linear hypothesis $\boldsymbol{f}^* = \Psi_m^\top \boldsymbol{\theta}_m^*$ with $\boldsymbol{\theta}_m^* \in I\!R^m, \Psi_m^\top \in I\!R^{n\times m}$ and for each m we assume that we can write the estimator $\widetilde{\boldsymbol{\theta}}_m$ as:

$$\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y} = (\Psi_m \Psi_m^\top)^{-1} \Psi_m \boldsymbol{Y}. \tag{4.1.2}$$

Except for the fact that we assume that we do not know Σ, we will otherwise use the same setting and the same notations as in Chapter 3. We recall the definitions (3.2.4) and (3.2.5) of $\boldsymbol{\xi}_{m,m^\circ}$ and $\boldsymbol{\xi}_m$. The joint distribution of $(\|\boldsymbol{\xi}_{m,m^\circ}\|)_{m^\circ \geq m; m, m^\circ \in \mathcal{M}}$ is key in the determination of the critical values. As we cannot sample it directly, we are going to use a bootstrapping procedure to approximate this distribution. We have to find some replacement

for the errors $\Sigma^{1/2}\boldsymbol{\varepsilon}$. The key idea, following the idea of the *wild* bootstrap explained above, is that use we use the reweighted residuals of a pilot estimator as a replacement for $\Sigma^{1/2}\boldsymbol{\varepsilon}$. We now introduce a probability measure $I\!P^\flat$ conditional on $\boldsymbol{Y}$ and for which we introduce the bootstrap error vector $\boldsymbol{\varepsilon}^\flat \sim \mathcal{N}(0, \mathbf{1}_n)$. We write $I\!E^\flat$ for the associated expectation. We need to presmooth $\boldsymbol{Y}$ to get residuals for which we have subtracted the main part of the signal. This pre-smoothing requires some minimal smoothness of the regression function, and this condition seems to be unavoidable if no information about the noise is given: otherwise one cannot distinguish between signal and noise. Below we suppose that a linear predictor $\widetilde{\boldsymbol{f}} = \Pi\boldsymbol{Y}$ is given where Π is a sub-projector in the space $I\!R^n$. For example, one can take $\Pi = \Psi_{m^\dagger}^\top \big(\Psi_{m^\dagger}\Psi_{m^\dagger}^\top\big)^{-1}\Psi_{m^\dagger}$ where $m^\dagger$ is a large model, e.g. the largest model $m_{\max}$ in our collection. Then one can compute the residuals $\breve{\boldsymbol{Y}} = \boldsymbol{Y} - \Pi\boldsymbol{Y}$ and use $\mathrm{diag}(\breve{\boldsymbol{Y}})$ as a replacement for the unknown $\Sigma^{1/2}$. The signs of the coefficients of $\breve{\boldsymbol{Y}}$ do not matter, as the coefficients will only appear in products $\breve{\boldsymbol{Y}} \cdot \boldsymbol{\varepsilon}^\flat$, which are invariant in distribution under changes of sign of the coefficients. We now define the following quantities after replacing $\Sigma^{1/2}\boldsymbol{\varepsilon}$ by $\mathrm{diag}(\breve{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat$:

$$\|\boldsymbol{\xi}^\flat_{m,m^\circ}\| \stackrel{\mathrm{def}}{=} \|\mathcal{K}_{m,m^\circ}\,\mathrm{diag}(\breve{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat\|,\ m^\circ, m \in \mathcal{M}. \tag{4.1.3}$$

The idea here is to subtract enough bias for our bootstrap statistics to be comparable to the real world statistics, but not subtract too much of the true error terms. We use the same arguments to get a bootstrap-equivalent of $\|\boldsymbol{\xi}_m\|$ in the form:

$$\|\boldsymbol{\xi}^\flat_m\| \stackrel{\mathrm{def}}{=} \|W(\Psi_m\Psi_m^\top)^{-1}\Psi_m\,\mathrm{diag}(\breve{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat\| = \|\mathcal{K}_m\,\mathrm{diag}(\breve{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat\|,$$

If we write $\mathbb{B}_m = \mathcal{K}_m\,\mathrm{diag}(\breve{\boldsymbol{Y}})$ and $\mathbb{B}_{m^\circ,m} = \mathcal{K}_{m,m^\circ}\,\mathrm{diag}(\breve{\boldsymbol{Y}})$, we can define analogously as in Eqn. (3.2.6) for $m^\circ, m \in \mathcal{M}$ the bootstrap *effective dimensions*:

$$\mathtt{p}^\flat_{m^\circ,m} \stackrel{\mathrm{def}}{=} \mathrm{tr}(\mathbb{B}^\top_{m^\circ,m}\mathbb{B}_{m^\circ,m}), \quad \mathtt{p}^\flat_m \stackrel{\mathrm{def}}{=} \mathrm{tr}(\mathbb{B}^\top_m\mathbb{B}_m)$$

and

$$\lambda^\flat_{m^\circ,m} \stackrel{\mathrm{def}}{=} \|\mathbb{B}^\top_{m^\circ,m}\mathbb{B}_{m^\circ,m}\|_{\mathrm{op}}, \quad \lambda^\flat_m \stackrel{\mathrm{def}}{=} \|\mathbb{B}^\top_m\mathbb{B}_m\|_{\mathrm{op}}$$

for the largest eigenvalues.

4.2 Calibrating the critical values

We recall that we can directly sample from $I\!P^\flat$ conditional on $\boldsymbol{Y}$. The algorithm for calibration in the known-variance case is now adjusted in the following way: We fix $\mathtt{x} > 0$ and

- simulate the tail functions $z^\flat_{m^\circ,m}(\cdot)$ of $\|\boldsymbol{\xi}^\flat_{m,m^\circ}\|$ in the bootstrap world

$$I\!P^\flat(\|\boldsymbol{\xi}^\flat_{m^\circ,m}\| \geq z^\flat_{m^\circ,m}(x)) = \exp(-x), \forall x \geq 0, \tag{4.2.1}$$

- choose for each m a $q^\flat_m \geq 0$ such that for $\mathtt{x}^\flat \stackrel{\text{def}}{=} \mathtt{x} + q^\flat_m$:

$$I\!P^\flat(\bigcup_{m^\circ=m}^{m_{\max}} \{\|\boldsymbol{\xi}^\flat_{m,m^\circ}\| \geq z^\flat_{m^\circ,m}(\mathtt{x}^\flat)\}) = \exp(-\mathtt{x}), \tag{4.2.2}$$

- and define $z^{\flat,+}_{m^\circ,m}(\mathtt{x}) \stackrel{\text{def}}{=} z^\flat_{m^\circ,m}(\mathtt{x}^\flat)$.

This gives us a way to replace $z^+_{m^\circ,m}(\mathtt{x})$ with a bootstrap equivalent. But as we do not know the effective dimension we are using in the bias bound either, we also need to replace $\mathtt{p}_{m,m^\circ}$ by a bootstrap-analog. We use straightforwardly:

$$\mathtt{p}^\flat_{m,m^\circ} \stackrel{\text{def}}{=} I\!E^\flat\|\boldsymbol{\xi}^\flat_{m,m^\circ}\|^2 = \mathrm{tr}(\mathbb{B}^\top_{m^\circ,m}\mathbb{B}_{m^\circ,m})$$

as a replacement. Finally, this means that we fix the critical values analogously to Eqn. (3.3.1) by:

$$\mathfrak{z}^\flat_{m,m^\circ} = z^\flat_{m,m^\circ}(\mathtt{x} + q^\flat_{m^\circ}) + \beta\sqrt{\mathtt{p}^\flat_{m,m^\circ}} \tag{4.2.3}$$

for some $\beta \geq 0$.

All the quantities we have calculated will be $I\!P$-random depending on the observed data $\boldsymbol{Y}$. This will make the analysis of the method more involved than in the known variance case.

The model selection step stays exactly the same as in the case of known variance, see (3.2.1). We choose the model by $\widehat{m}$:

$$\widehat{m} \stackrel{\text{def}}{=} \mathrm{argmin}\big\{m^\circ \in \mathcal{M}\colon \max_{m \in \mathcal{M}^+(m^\circ)} \big\{\mathbb{T}_{m,m^\circ} - \mathfrak{z}^\flat_{m,m^\circ}\big\} \leq 0\big\}.$$

Now we are going to study the theoretical properties of the method under suitable conditions in comparison to the case of known variance.

4.3 Theoretical properties

In this section, we want to show that the same results, which held for the probabilistic loss for the known variance case, continue to hold up to small correction terms in the bootstrap version of the method. In the following, we define the quantities that will govern how well the bootstrap method will perform. We write $\Psi_{m,i}$ for the i th column of Ψ_m.

We measure the **design regularity** by the value δ_Ψ

$$\delta_\Psi \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}} \max_{1 \le i \le n} \sigma_i \|S_m^{-1/2} \Psi_{m,i}\|. \tag{4.3.1}$$

where $S_m = \Psi_m \Sigma \Psi_m^\top$.

The **presmoothing bias** for the presmoothing operator Π is measured by the vector

$$\boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi \boldsymbol{f}^*).$$

It is the approximation bias of $\boldsymbol{f}^*$ by $\Pi \boldsymbol{f}^*$ weighted by the standard deviations of the noise terms.

We also measure the **presmoothing stochastic noise** in terms of the covariance matrix $\mathrm{Var}(\check{\boldsymbol{\varepsilon}})$ of the smoothed noise $\check{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} - \Pi_\Sigma \boldsymbol{\varepsilon}$, where $\Pi_\Sigma \stackrel{\text{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$. Namely, this matrix is assumed to be sufficiently close to the unit matrix $\mathbf{1}_n$, in particular, its diagonal elements should be close to one. This is measured by the operator norm of $\mathrm{Var}(\check{\boldsymbol{\varepsilon}}) - \mathbf{1}_n$ and by the deviation of the individual variances $I\!\!E \check{\varepsilon}_i^2$ from one. We also need a control on the

maximal variance of the individual variances of $\breve{\varepsilon}_i - \varepsilon_i$. We write this as:

$$\begin{aligned} c_1 &\stackrel{\text{def}}{=} \| \operatorname{Var}(\breve{\boldsymbol{\varepsilon}}) - \mathbf{1}_n \|_{\text{op}}, \\ \delta_1 &\stackrel{\text{def}}{=} \max_{1 \le i \le n} \sqrt{\operatorname{Var}(\breve{\varepsilon}_i - \varepsilon_i)}, \\ \delta_2 &\stackrel{\text{def}}{=} \max_{1 \le i \le n} |I\!E \breve{\varepsilon}_i^2 - 1|^{1/2}. \end{aligned}$$

In particular, in the case of homogeneous errors $\Sigma = \sigma^2 \mathbf{1}_n$ and assuming the presmoothing operator Π to be a p-dimensional projector of the form $\Pi = \Psi_{m^\dagger}^\top \bigl(\Psi_{m^\dagger} \Psi_{m^\dagger}^\top\bigr)^{-1} \Psi_{m^\dagger}$ for some model $m^\dagger \in \mathcal{M}$. We have

$$\operatorname{Var}(\breve{\boldsymbol{\varepsilon}}) = (\mathbf{1}_n - \Pi)^2 = \mathbf{1}_n - \Pi \le \mathbf{1}_n,$$

and

$$\begin{aligned} c_1 &= \| \operatorname{Var}(\breve{\boldsymbol{\varepsilon}}) - \mathbf{1}_n \|_{\text{op}} = \|\Pi\|_{\text{op}} = 1, \\ \delta_1^2 &= \max_{1 \le i \le n} \operatorname{Var}(\breve{\varepsilon}_i - \varepsilon_i) = \max_{1 \le i \le n} \Psi_{m^\dagger, i}^\top \bigl(\Psi_{m^\dagger} \Psi_{m^\dagger}^\top\bigr)^{-1} \Psi_{m^\dagger, i} \\ \delta_2^2 &= \max_{1 \le i \le n} |I\!E \breve{\varepsilon}_i^2 - 1| = \max_{1 \le i \le n} \Psi_{m^\dagger, i}^\top \bigl(\Psi_{m^\dagger} \Psi_{m^\dagger}^\top\bigr)^{-1} \Psi_{m^\dagger, i}, \end{aligned}$$

and δ_1, δ_2 will usually be of order $\sqrt{\frac{p}{n}}$, if $\Psi_{m^\dagger}$ satisfies some Lindeberg-type condition similar to (4.3.1). In the following theorem, we will show that when using the bootstrap calibrated bounds $z^\flat_{m,m^\circ}(\mathtt{x} + q^\flat_{m^\circ})$ instead of $z_{m,m^\circ}(\mathtt{x} + q_{m^\circ})$, we will get almost the same probability statements for the stochastic noise terms $\boldsymbol{\xi}_{m,m^\circ}$. For the statement of the theorem, we assume that we can write our models as the projection of a larger model onto a smaller feature set. We write $\Psi = \Psi_{m_{\max}} \in I\!R^{p \times n}$ for the design matrix of the largest model $m_{\max} \in \mathcal{M}$ with feature dimension p. We assume that the Ψ_m from (4.1.2) can be written as projections of the largest model onto a smaller feature set:

$$\Psi_m = \Pi_m \Psi$$

for a projector Π_m.

Theorem 4.3.1. *Let $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ be a Gaussian vector in $I\!R^n$ with independent components, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{f}^*, \Sigma)$ for $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Let*

the design matrix $\Psi \in I\!R^{p\times n}$ *of the largest model in* $\mathcal{M}$, *be such that* $S = \Psi\Sigma\Psi^\top \in I\!R^{p\times p}$ *is invertible. For a given presmoothing operator* $\Pi\colon I\!R^n \to I\!R^n$ *let the values* $z^\flat_{m,m^\circ}(\mathtt{x})$ *and* $q^\flat_{m^\circ}$ *for all* $m > m^\circ$ *be defined by* (4.2.2) *and* (4.2.1). *Then it holds*

$$I\!P\left(\max_{m>m^\circ}\left\{\|\boldsymbol{\xi}_{m,m^\circ}\| - z^\flat_{m,m^\circ}(\mathtt{x}+q^\flat_{m^\circ})\right\} \ge 0\right) \le 8\exp(-\mathtt{x}) + \Delta_{\boldsymbol{\xi}}(\mathtt{x}).$$

with

$$\begin{aligned}\Delta_{\boldsymbol{\xi}}(\mathtt{x}) \stackrel{\text{def}}{=} p^{1/2}&(4\delta_1^2\mathtt{x}_n + 4\sqrt{2}\delta_1\mathtt{x}_n + 4\sqrt{2}\|\boldsymbol{B}\|_\infty\delta_1\sqrt{\mathtt{x}_n} + 2\delta_\Psi\sqrt{\mathtt{x}+\log(2p)}\\ &+2\delta_\Psi^2(\mathtt{x}+\log(2p)) + \|\boldsymbol{B}\|_\infty^2 + \delta_\Psi^2\|\boldsymbol{B}\|\sqrt{2\mathtt{x}},\end{aligned}$$

if $\Delta_{\boldsymbol{\xi}}/p^{1/2} \le 1/2$.

To give a more readable version of the bound, we make the assumption, that we can bound δ_Ψ, δ_1 by some common δ and $2p \le n$, then we can get a simpler bound:

$$\Delta_{\boldsymbol{\xi}}(\mathtt{x}) \le \mathtt{C}\mathtt{x}_n p^{1/2}\left(\delta + \|\boldsymbol{B}\|_\infty\delta + \|\boldsymbol{B}\|_\infty^2 + \|\boldsymbol{B}\|\delta^2\right)$$

for some numerical constant $\mathtt{C} > 0$.

The SmA procedure also involves the values $\mathtt{p}_{m,m^\circ}$ which are unknown and depend on the noise structure Σ. The next result shows that the bootstrap counterparts $\mathtt{p}^\flat_{m,m^\circ}$ can be well used in place of $\mathtt{p}_{m,m^\circ}$.

Theorem 4.3.2. *Assume the conditions of Theorem 4.3.1. Then it holds for* $\mathtt{p}^\flat_{m,m^\circ} = I\!E^\flat\|\boldsymbol{\xi}^\flat_{m,m^\circ}\|^2$ *:*

$$I\!P\left(\max_{m>m^\circ, m,m^\circ\in\mathcal{M}}\left|\frac{\mathtt{p}^\flat_{m,m^\circ}}{\mathtt{p}_{m,m^\circ}} - 1\right| > \Delta_{\mathtt{p}}(\mathtt{x})\right) \le 3\exp(-\mathtt{x}),$$

where

$$\Delta_{\mathtt{p}}(\mathtt{x}) \stackrel{\text{def}}{=} 4\mathtt{x}_{\mathcal{M}}^{1/2}\delta_\Psi + 2\,\mathtt{x}_{\mathcal{M}}\,\delta_\Psi^2 + \|\boldsymbol{B}\|_\infty^2 + 4\,\mathtt{x}_{\mathcal{M}}^{1/2}\,\delta_\Psi^2\,\|\boldsymbol{B}\| + \delta_2$$

with $\mathtt{x}_{\mathcal{M}} = \mathtt{x} + 2\log(|\mathcal{M}|)$.

Again we are giving a simplified bound under the assumption that some δ bounds δ_2 and δ_Ψ from above and that $\mathtt{x} \geq 1$.

$$\Delta_{\mathtt{p}} \leq \mathtt{C}\mathtt{x}_{\mathcal{M}} \left(\delta + \delta^2 \|\boldsymbol{B}\| + \|\boldsymbol{B}\|_\infty^2\right)$$

for some numerical constant $\mathtt{C} > 0$. Finally, we show that the bounds we used in Theorem 3.4.2 in the section to bound the payment for adaptation are the same as for the bootstrap-case up to some correction term:

Proposition 4.3.3. *Under the conditions of Theorem 4.3.1*

$$\mathbb{P}\left(\max_{m>m^\circ}\left\{\mathfrak{z}^\flat_{m,m^\circ} - K_{\mathfrak{z}}(\mathtt{x})\left((1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} - \sqrt{2\lambda_{m,m^\circ}\mathtt{x}_{\mathcal{M}}}\right)\right\} \geq 0\right)$$
$$\leq 3\exp(-\mathtt{x}) + \Delta_{\boldsymbol{\xi}}(\mathtt{x}),$$

where

$$K_{\mathfrak{z}}(\mathtt{x}) = \max\{\sqrt{1+\Delta_{\boldsymbol{\xi}}(\mathtt{x})}, \sqrt{1+\Delta_{\mathtt{p}}(\mathtt{x})}\}$$

with $\Delta_{\boldsymbol{\xi}}, \Delta_{\mathtt{p}}$ *from Theorems 4.3.1 and 4.3.2 and again* $\mathtt{x}_{\mathcal{M}} = \mathtt{x} + 2\log(|\mathcal{M}|)$.

The above results allow to extend all the oracle bounds for probabilistic loss of Chapter 3 with the obvious corrections of the error probability. We will give one example of such an extension in Theorem 4.3.5 further below.

But first we discuss the sense of the required conditions for bootstrap validity. In typical situations, we have $\delta = \max\{\delta_\Psi, \delta_2, \delta_1\} \leq \mathtt{C}\sqrt{\frac{p}{n}}$. One can see that the bootstrap approximation is accurate if the values $\Delta_{\boldsymbol{\xi}}$ and $\Delta_{\mathtt{p}}$ are small. This requires that the values $\delta^2 p$, $\|\boldsymbol{B}\|_\infty^4 p$, and $\delta^2\|\boldsymbol{B}\|$ are sufficiently small. It is easy to see that the last term is smaller in order than the others. We have

$$\delta^2 p = \mathcal{O}(p^2/n).$$

Further, the bias component does not damage the bootstrap validity result if $\|\boldsymbol{B}\|_\infty^4 p$ is a small value. If $\boldsymbol{f}^*$ is Hölder-smooth with the parameter s, that is, if

$$\|\boldsymbol{B}\|_\infty \leq \mathtt{C}p^{-s}, \tag{4.3.2}$$

then the bootstrap procedure is justified for $s > 1/4$ if $p = p_n \to \infty$ as $n \to \infty$ and $p^2 \log(n)/n$ goes to zero. We state one asymptotic result of this sort.

Corollary 4.3.4. *Assume that* $\delta = \max\{\delta_\Psi, \delta_2, \delta_1\} \leq \mathtt{C}\sqrt{\frac{p}{n}}$. *Let also* $p = p_n$ *satisfy* $p_n^2 \log(n)/n \to 0$ *as* $n \to \infty$, *and* (4.3.2) *hold for* $s > 1/4$. *Then the results of Theorem 4.3.1, 4.3.2 and Proposition 4.3.3 apply with a small value* $\Delta_{\boldsymbol{\xi}} = \Delta_{\boldsymbol{\xi},n} \to 0$ *as* $n \to \infty$.

We now give a bootstrap version of Theorem 3.4.1. We have to change the definition of m^* slightly for this. We define:

$$m^* \stackrel{\text{def}}{=} \min\Big\{ m^\circ \in \mathcal{M} \colon \max_{m > m^\circ} \big\{ \|\boldsymbol{b}_{m,m^\circ}\|^2 - \beta^2 (1 + \Delta_{\mathtt{p}})\, \mathtt{p}_{m,m^\circ} \big\} \leq 0 \Big\}. \tag{4.3.3}$$

If we assume to be in a case where $\Delta_{\mathtt{p}}$ is small, this means that we slightly change the value of β. For $\Delta_{\mathtt{p}}$ going to zero for n going to infinity this definition coincides asymptotically with our original definition . We are now ready to give the following probabilistic oracle result.

Theorem 4.3.5. *Assume the conditions of Theorem 4.3.1. Given* $\mathtt{x}$ *and* β, *let the critical values for the SmA-method be given by* $\mathfrak{z}^\flat_{m,m^\circ}$ *from* (4.2.3) *and let* $m^* \in \mathcal{M}$ *satisfy* (4.3.3). *Then the SmA estimator* $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ *satisfies the following bound:*

$$I\!P\Big(\big\| \widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^*} \big\| > \bar{\mathfrak{z}}^\flat_{m^*} \Big) \leq 11 \exp(-\mathtt{x}) + \Delta_{\boldsymbol{\xi}}(\mathtt{x}), \tag{4.3.4}$$

where $\bar{\mathfrak{z}}^\flat_{m^*}$ *is defined as*

$$\bar{\mathfrak{z}}^\flat_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^+(m^*)} \mathfrak{z}^\flat_{m^*,m}\,.$$

This implies the probabilistic oracle bound: with probability at least $1 - 11\exp(-\mathtt{x}) - \Delta_{\boldsymbol{\xi}}(\mathtt{x})$

$$\big\| \widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^* \big\| \leq \big\| \widetilde{\boldsymbol{\phi}}_{m^*} - \boldsymbol{\phi}^* \big\| + \bar{\mathfrak{z}}^\flat_{m^*}. \tag{4.3.5}$$

In the next section, we are going to present some simulations for the method.

4.4 Simulations

This section illustrates the performance of the proposed procedure by means of simulated examples. We consider a regression problem for an unknown univariate function on $[0,1]$ with unknown heteroscedastic noise. The aim is to compare the bootstrap-calibrated procedure with the SmA procedure for the known noise and with the oracle estimator. We use a model $m^\dagger$ to accomplish the presmoothing: $\Pi = \Psi_{m^\dagger}^\top \left(\Psi_{m^\dagger}\Psi_{m^\dagger}^\top\right)^{-1}\Psi_{m^\dagger}$. We also check the sensitivity of the method to the choice of the model used for the presmoothing.

We use a uniform design on $[0,1] \subset I\!R$ and the Fourier basis $\{\psi_j(x)\}_{j=1}^\infty$ for approximation of the regression function f which is modeled in the form

$$f(x) = \sum_{j=1}^{p} c_j \psi_j(x),$$

where the $(c_j)_{1\le j\le p}$ were chosen randomly with

$$c_j = \begin{cases} \gamma_j, 1 \le j \le 10, \\ \gamma_j/(j-10)^2, 11 \le j \le 200, \end{cases}$$

and γ_j are i.i.d. standard normal. The noise intensity grows from low to high as x increases to one. We simulate the bootstrap tail functions with $n_{\text{sim-bs}} = 10^3$ simulated samples and the Monte-Carlo tail functions (under the assumption of a known noise structure) with $n_{\text{sim-mc}} = 10^3$ simulated samples. In this sieve setup, we will use m to denote the model itself and the model dimension. The maximal model dimension is chosen as $m_{\max} = 34$ and we choose $m^\dagger = 20$. The calibration is run with $n_{\text{sim-calib}} = 10^3$ and $\mathtt{x} = 2$, $\beta = 1$.

We start by considering examples for $W = \Psi_n^\top$, i.e. the estimation of the whole function vector with prediction loss. In Figure 4.4.1, one can see three examples with different intensities of the noise term comparing the bootstrap-method to the oracle estimator and the known-variance SmA-Method. Figure 4.4.2 illustrates the dependence of the choice of the estimated dimension on our calibration dimension $m^\dagger$ and the sample size n. We see that in the specific example we are considering, the sensitivity of

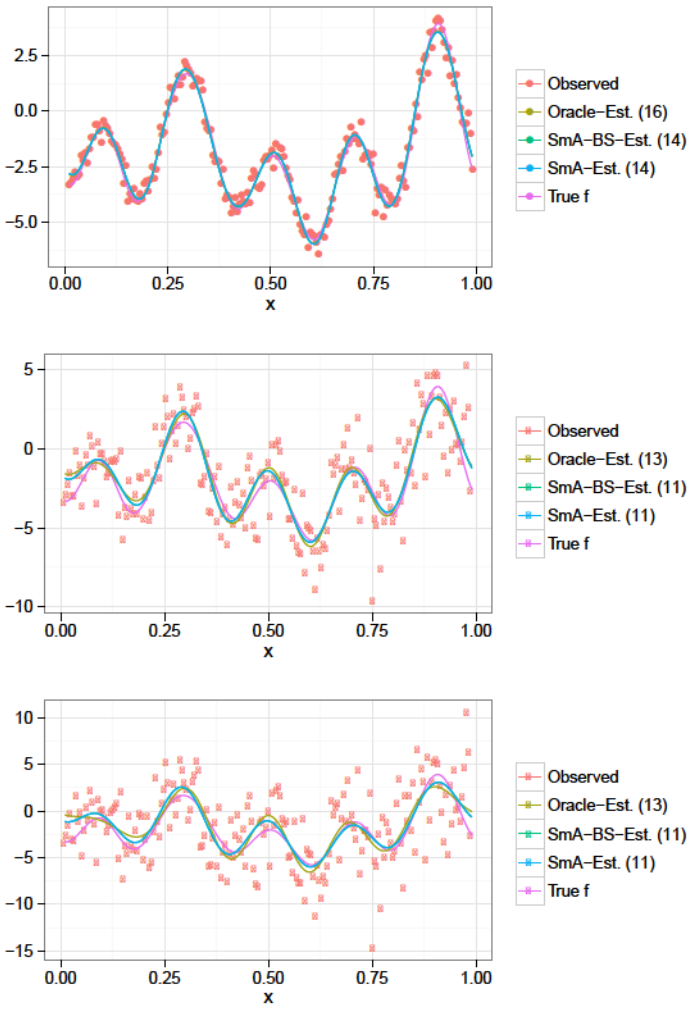

Figure 4.4.1: True functions and observed values plotted with oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.) for 3 different functions with different noise structure going from low noise to high noise. The numbers in parentheses indicate the chosen model dimension.

the chosen dimension $\widehat{m}$ on $m^\dagger$ decreases very fast. In the cases $n = 200$ and $n = 100$, we have no variation in the choice of $\widehat{m}$ with respect to $m^\dagger$. The oracles are respectively $m^* = 12$ for $n = 100, 200$ and $m^* = 10$ for $n = 50$. We also want to compare the true quantiles and their bootstrap substitute. Figure 4.4.3 plots the ratios of quantiles for all possible comparisons (m_1, m_2) for the same function as before. Here we see that there is, as one would expect, still significant variation in the quantile ratios for small differences $|m_1 - m_2|$. Nonetheless the method works very well as seen in Fig. 4.4.2, but the variability in the ratios implies the possibility to perhaps stabilize the procedure even more by introducing some smoothing scheme for the quantiles.

Figure 4.4.4 again demonstrates the dependence of the ratios on $m^\dagger$. One notices that the ratio is varying very slowly above $m^* = 12$.

We also give the results on the simulation of $n_{\text{hist}} = 100$ repeated applications of the method to the same true underlying function observed with different realizations of the errors in Figure 4.4.5. We observe that the known-variance and the bootstrap version behave very similar in their choices of a model. The bootstrap method only shows slightly more variation than the Monte-Carlo method. Up to now the implementation of the calibration algorithm, which is implemented in R is slow, which explains the rather small number of simulations we conducted. To make the algorithm usable for real applications, one would need to design a faster implementation in some faster programming language.

The case of the estimation of the first derivative is similar. We now choose

$$W = (\psi_j^{(k)}(x_i))_{i,j})_{1 \le i,j \le n}$$

and otherwise stay in the same setup as in the example before. Figure 4.4.6 shows an example of the application of the method with the new W. One can note that the SmA-procedure does a decent job of mimicking the oracle.

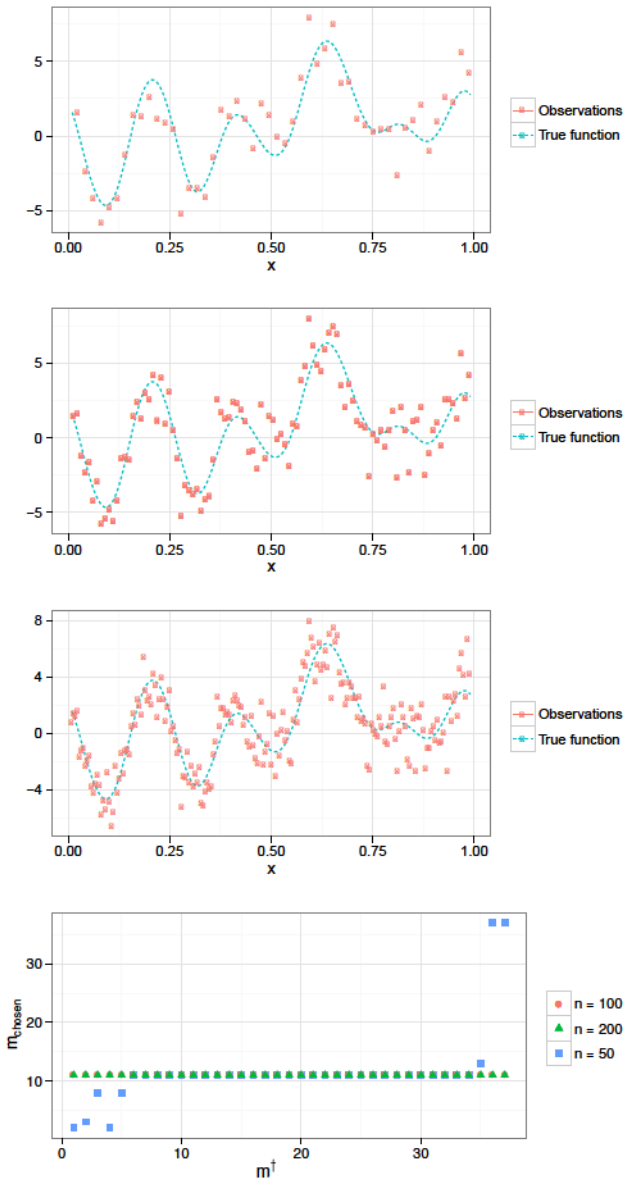

Figure 4.4.2: The first three plots show an exemplary function with $n = 50, 100, 200$ observations. The last plot shows the $\widehat{m}$ chosen by the Bootstrap-SmA-Method as a function of the calibration dimension $m^{\dagger}$ and the number of observations.

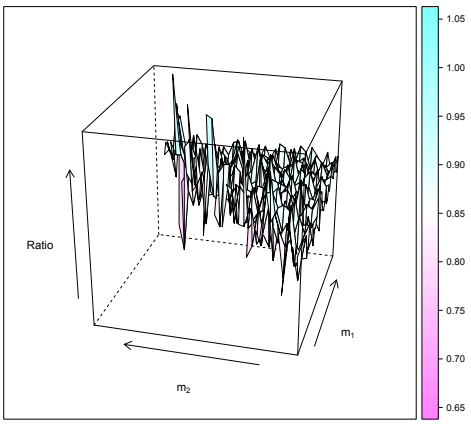

Figure 4.4.3: Ratio of quantiles $|\mathfrak{z}^{\flat}_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}|^2$ for $m^{\dagger} = 20$ and $n = 200$ with the data and true function as in Fig. 4.4.2.

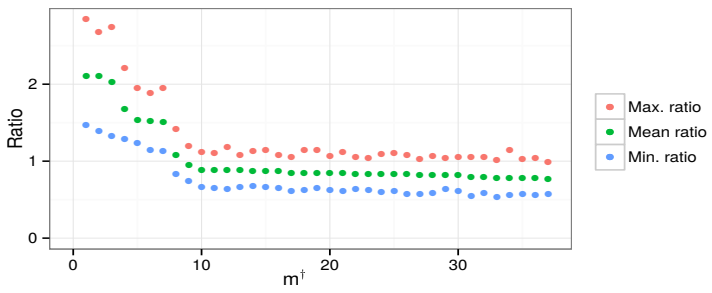

Figure 4.4.4: Maximal, minimal and mean ratio of the bootstrap and theoretical tail functions at $x = 2$, $|\mathfrak{z}^{\flat}_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}|^2, m_1, m_2 \in \mathcal{M}$ as a function of $m^{\dagger}$.

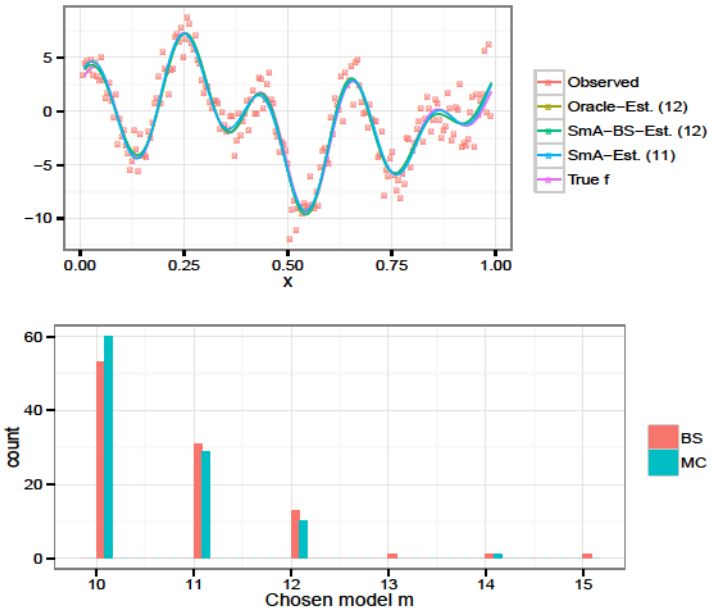

Figure 4.4.5: In the upper plot, the true function and observed values are plotted for one realization together with the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The numbers in parentheses indicate the chosen model dimension. In the lower plot, histograms for the selected model are given for the bootstrap (BS) and the known-variance method (MC) for repeated observations of the same underlying function with a simulation size $n_{\text{hist}} = 100$.

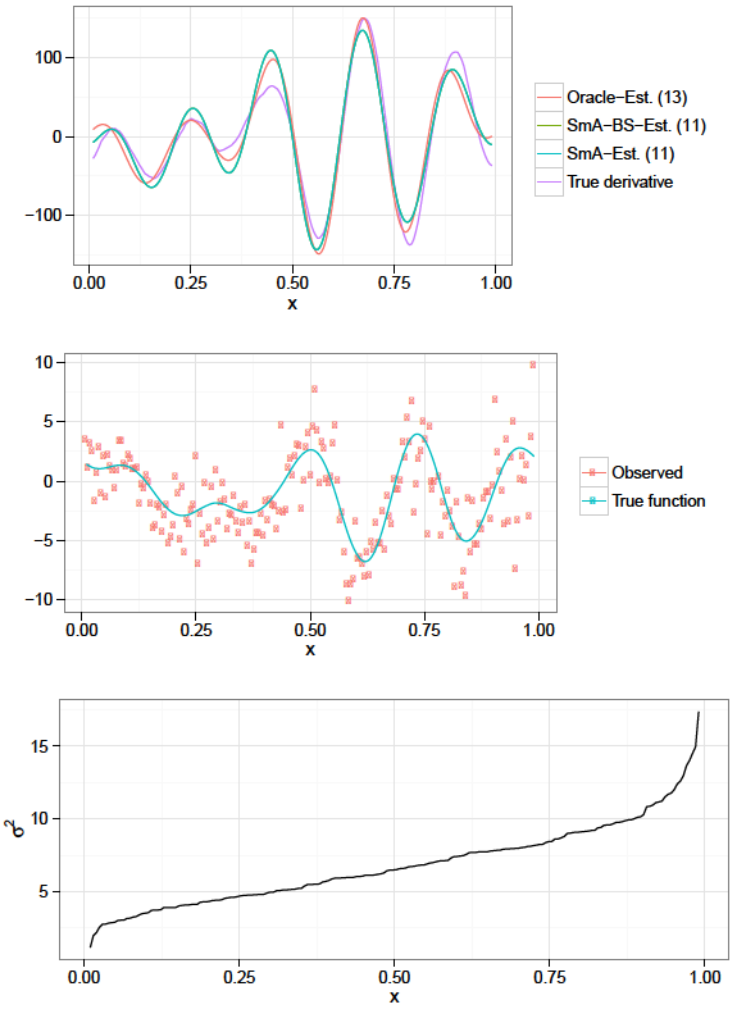

Figure 4.4.6: In the uppermost plot, we compare the true derivative, the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The next plot shows the true function and the observations and in the last plot one can find the standard deviation of the errors as a function of the design points x.

4.5 Proofs

4.5.1 Proof of Theorem 4.3.1

As in the statement of Theorem 4.3.1, we will write m° for a fixed model, which we compare to all larger models $m \in \mathcal{M}^+(m^\circ)$ Below we write Ψ in place of $\Psi_{m_{\max}}$, where $m_{\max}$ is the largest model in the collection. By p we denote the corresponding parameter dimension, that is, Ψ is a $p \times n$ matrix. Furthermore, the feature matrix Ψ_m can be written as the product $\Psi_m = \Pi_m \Psi$, where Π_m is the projector on the subspace of the feature space spanned by the features from the m-model: $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m$. This allows to represent each estimator $\widetilde{\phi}_m$ in the form

$$\widetilde{\phi}_m = W \widetilde{\boldsymbol{\theta}}_m = W \mathcal{S}_m \boldsymbol{Y} = W \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m \boldsymbol{Y} = \mathcal{T}_m \Psi \boldsymbol{Y}$$

with

$$\mathcal{T}_m \stackrel{\text{def}}{=} W \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Pi_m .$$

This implies the following representation of the stochastic components $\boldsymbol{\xi}_{m,m^\circ}$:

$$\begin{aligned} \boldsymbol{\xi}_{m,m^\circ} &= \mathcal{T}_{m,m^\circ} \Psi \boldsymbol{\varepsilon} = \mathcal{T}_{m,m^\circ} \boldsymbol{\zeta}, \\ \mathcal{T}_{m,m^\circ} &\stackrel{\text{def}}{=} \mathcal{T}_m - \mathcal{T}_{m^\circ} , \end{aligned}$$

where $\boldsymbol{\zeta} = \Psi \Sigma^{1/2} \boldsymbol{\varepsilon}$. One sees that each stochastic vector $\boldsymbol{\xi}_{m,m^\circ}$ is a linear transformation of the vector $\boldsymbol{\zeta}$. A similar representation holds true in the bootstrap world:

$$\begin{aligned} \boldsymbol{\xi}^\flat_{m,m^\circ} &= \mathcal{T}_{m,m^\circ} \Psi \operatorname{diag}(\check{\boldsymbol{Y}}) \boldsymbol{\varepsilon}^\flat = \mathcal{T}_{m,m^\circ} \boldsymbol{\zeta}^\flat, \\ \boldsymbol{\zeta}^\flat &\stackrel{\text{def}}{=} \Psi \operatorname{diag}(\check{\boldsymbol{Y}}) \boldsymbol{\varepsilon}^\flat. \end{aligned}$$

Here the original errors $\Sigma^{1/2} \boldsymbol{\varepsilon}$ are replaced by their bootstrap counterparts $\operatorname{diag}(\check{\boldsymbol{Y}}) \boldsymbol{\varepsilon}^\flat$ as explained in Section 3.1. Normality of the errors $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), 1 \le i \le n,$ implies that $\boldsymbol{\zeta} = \Psi \Sigma^{1/2} \boldsymbol{\varepsilon}$ is also normal zero mean:

$$\boldsymbol{\zeta} \sim \mathcal{N}(0, S), \qquad S \stackrel{\text{def}}{=} \Psi \Sigma \Psi^\top .$$

Similarly, we can use standard normality of the bootstrap errors $\boldsymbol{\varepsilon}^\flat$. Given the data $\boldsymbol{Y}$, the vector $\boldsymbol{\zeta}^\flat$ is conditionally normal zero mean with the conditional variance

$$S^\flat \stackrel{\mathrm{def}}{=} \operatorname{Var}^\flat(\boldsymbol{\zeta}^\flat) = \Psi \operatorname{diag}\bigl(\breve{Y}_1^2, \ldots, \breve{Y}_n^2\bigr)\Psi^\top = \Psi \operatorname{diag}\bigl(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}}\bigr)\Psi^\top.$$

Therefore, the problem is almost reduced to the problem of comparing the the two p-dimensional Gaussian distributions of $\boldsymbol{\zeta} = \Psi\Sigma^{1/2}\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}^\flat = \Psi \operatorname{diag}(\breve{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat$ given $\boldsymbol{Y}$ in total variation distance. One has to be careful though. To show our result, we want to bound the difference between

$$I\!P\left(\max_{m>m^\circ}\Bigl\{\|\boldsymbol{\xi}_{m,m^\circ}\| - z^\flat_{m,m^\circ}(\mathtt{x} + q^\flat_{m^\circ})\Bigr\} \geq 0\right)$$

and

$$I\!P\left(\max_{m>m^\circ}\Bigl\{\|\boldsymbol{\xi}_{m,m^\circ}\| - z_{m,m^\circ}(\mathtt{x} + q_{m^\circ})\Bigr\} \geq 0\right),$$

but $z^\flat_{m,m^\circ}(\mathtt{x}+q^\flat_{m^\circ})$ depending on $\boldsymbol{Y}$ is no deterministic set in the $I\!P$-world and therefore, we cannot directly compare these two quantities by the total variation distance. We must make a detour of sandwiching the bootstrap $I\!P^\flat$ with high probability between two deterministic measures $I\!P^+, I\!P^-$ independent of $\boldsymbol{Y}$.

$$\mathcal{B} \stackrel{\mathrm{def}}{=} S^{-1/2}\bigl(S^\flat - S\bigr)S^{-1/2}.$$

We define the matrix $\mathcal{U} \stackrel{\mathrm{def}}{=} S^{-1/2}\Psi\Sigma^{1/2} \in I\!R^{p\times n}$. It holds

$$\mathcal{U}\mathcal{U}^\top = \mathbf{1}_p.$$

We will use the decomposition

$$\begin{aligned}
\Sigma^{-1/2}\breve{\boldsymbol{Y}} &= \Sigma^{-1/2}(\boldsymbol{Y} - \Pi\boldsymbol{Y}) \\
&= \Sigma^{-1/2}(\Sigma^{1/2}\boldsymbol{\varepsilon} - \Pi\Sigma^{1/2}\boldsymbol{\varepsilon}) + \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*) \\
&= \breve{\boldsymbol{\varepsilon}} + \boldsymbol{B}
\end{aligned}$$

with

$$\check{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} - \Sigma^{-1/2} \Pi \Sigma^{1/2} \boldsymbol{\varepsilon},$$
$$\boldsymbol{B} = \Sigma^{-1/2} \bigl(\boldsymbol{f}^* - \Pi \boldsymbol{f}^* \bigr) \tag{4.5.1}$$

and we also define

$$\boldsymbol{Y}' \stackrel{\text{def}}{=} (\boldsymbol{f}^* - \Pi \boldsymbol{f}^*) + \Sigma^{1/2} \boldsymbol{\varepsilon} \tag{4.5.2}$$

as a modification of $\boldsymbol{Y}$, where we neglect the influence of presmoothing on the stochastic noise term. The matrix $\mathcal{B}$ can now be represented as

$$\mathcal{B} = \mathcal{U} \operatorname{diag}\bigl\{ (\check{\boldsymbol{\varepsilon}} + \boldsymbol{B}) \cdot (\check{\boldsymbol{\varepsilon}} + \boldsymbol{B}) - \mathbf{1}_n \bigr\} \mathcal{U}^\top.$$

Now we introduce the sets

$$A(\mathtt{x}) \stackrel{\text{def}}{=} \bigcup_{m \in \mathcal{M}^+(m^\circ)} \{ \boldsymbol{u} \colon \| \mathcal{T}_{m,m^\circ} \boldsymbol{u} \| \le z_{m,m^\circ}(\mathtt{x}) \} \tag{4.5.3}$$

$$A^\flat(\mathtt{x}) \stackrel{\text{def}}{=} \bigcup_{m \in \mathcal{M}^+(m^\circ)} \Bigl\{ \boldsymbol{u} \colon \| \mathcal{T}_{m,m^\circ} \boldsymbol{u} \| \le z^\flat_{m,m^\circ}(\mathtt{x}) \Bigr\}$$

By construction, we have for $\mathtt{x}^\flat \stackrel{\text{def}}{=} \mathtt{x} + q^\flat_{m^\circ}$

$$I\!P^\flat \Bigl(A^\flat(\mathtt{x}^\flat) \Bigr) = 1 - \exp(-\mathtt{x}) \tag{4.5.4}$$

Below we use the operator norm for quantifying the difference between S and $S^\flat$: for now we assume that we are in the situation:

$$\| S^{-1/2} \, S^\flat \, S^{-1/2} - \mathbf{1}_p \|_{\mathrm{op}} \le \delta. \tag{4.5.5}$$

We recall that we want to answer the question whether the random multiplicity correction based on (4.5.4) does a good job under $I\!P$. This question leads to the analysis of $I\!P\bigl(A^\flat(\mathtt{x}^\flat)\bigr)$. Our aim now is to get a bound on the difference :

$$I\!P\bigl(A^\flat(\mathtt{x}^\flat)\bigr) - (1 - \exp(-\mathtt{x}))$$

conditionally on (4.5.5).

Theorem 4.5.1. *Let* $S^\flat$ *and* S *from above satisfy* (4.5.5) *for a* $\delta < 1/2$. *Then it holds with* $\Delta = \delta\sqrt{p}$

$$\bigl|I\!P\bigl(A^\flat(\mathtt{x}^\flat)\bigr) - (1 - \exp(-\mathtt{x}))\bigr| \le \Delta. \tag{4.5.6}$$

Proof. The key property of $I\!P^\flat = \mathcal{N}(0, S^\flat)$ is that the random matrix $S^\flat$ concentrates around the deterministic matrix S by (4.5.5). Below we use this property in the bracketing form:

$$S^- \le S^\flat \le S^+$$

with

$$S^- \stackrel{\text{def}}{=} (1-\delta)S, \qquad S^+ \stackrel{\text{def}}{=} (1+\delta)S, \qquad S^+ - S^- = 2\delta S. \tag{4.5.7}$$

In other words, the random matrix $S^\flat$ can be sandwiched between two deterministic matrices S^- and S^+ on the set where (4.5.5) holds.

For the proof of (4.5.6) we use the following well-known property of the Gaussian distribution.

Lemma 4.5.2. *Let* $I\!P_1 \sim \mathcal{N}(0, V_1)$ *and* $I\!P_2 \sim \mathcal{N}(0, V_2)$ *with* $V_1 \le V_2$. *Then for any centrally symmetric star-shaped set* A, *it holds*

$$I\!P_1(A) \ge I\!P_2(A).$$

Proof. The statement is trivial in the univariate case and the general case is obtained by integration over A in polar coordinates. □

We now introduce two Gaussian measures $I\!P^- = \mathcal{N}(0, S^-)$ and $I\!P^+ = \mathcal{N}(0, S^+)$; see (4.5.7). Let $z^-_{m,m^\circ}(\mathtt{x})$ and $z^+_{m,m^\circ}(\mathtt{x})$ be the corresponding tail functions, and $A^-(\mathtt{x})$ and $A^+(\mathtt{x})$ - the corresponding sets of the type constructed in Equation (4.5.3). The identities (4.5.7) yield for each $\mathtt{x}$ the relation

$$I\!P^+\bigl(A^+(\mathtt{x})\bigr) = I\!P^-\bigl(A^-(\mathtt{x})\bigr). \tag{4.5.8}$$

Lemma 4.5.2 implies by (4.5.7) for any $\mathtt{x}$

$$I\!P^+(A(\mathtt{x})) \le I\!P^\flat(A(\mathtt{x})) \le I\!P^-(A(\mathtt{x})). \tag{4.5.9}$$

The key step of the proof is given by the next lemma where we sandwich the random set $A^\flat(\mathtt{x}^\flat)$ in two specially constructed deterministic sets.

Lemma 4.5.3. *Define the deterministic values* $\mathtt{x}^-$ *and* $\mathtt{x}^+$ *by the equations*

$$\begin{aligned} I\!P^+\big(A^-(\mathtt{x}^+)\big) &= 1-\exp(-\mathtt{x}), \\ I\!P^-\big(A^+(\mathtt{x}^-)\big) &= 1-\exp(-\mathtt{x}). \end{aligned} \tag{4.5.10}$$

Then

$$\begin{aligned} \mathtt{x}^- \le \quad \mathtt{x}^\flat \quad &\le \mathtt{x}^+ \\ A^-(\mathtt{x}^-) \subseteq A^\flat(\mathtt{x}^\flat) &\subseteq A^+(\mathtt{x}^+). \end{aligned} \tag{4.5.11}$$

Proof. Before proving the result, we remark the cross-combination of the sets A^- with the probability measure $I\!P^+$ and the other way around. This crossing of the sets and measures, will be the main ingredient for the result. By Lemma 4.5.2 the following inequalities and inclusions hold true for any $\mathtt{x}$:

$$\begin{aligned} z^-_{m,m^\circ}(\mathtt{x}) \le z^\flat_{m,m^\circ}(\mathtt{x}) &\le z^+_{m,m^\circ}(\mathtt{x}), \\ A^-(\mathtt{x}) \subseteq \quad A^\flat(\mathtt{x}) \quad &\subseteq A^+(\mathtt{x}). \end{aligned} \tag{4.5.12}$$

Now by definition (4.5.10) in view of (4.5.9) and (4.5.12):

$$\begin{aligned} I\!P^\flat\big(A^\flat(\mathtt{x}^+)\big) \ge I\!P^+\big(A^\flat(\mathtt{x}^+)\big) \ge I\!P^+\big(A^-(\mathtt{x}^+)\big) &= 1-\exp(-\mathtt{x}), \\ I\!P^\flat\big(A^\flat(\mathtt{x}^-)\big) \le I\!P^-\big(A^\flat(\mathtt{x}^-)\big) \le I\!P^-\big(A^+(\mathtt{x}^-)\big) &= 1-\exp(-\mathtt{x}). \end{aligned}$$

This yields by monotonicity of $I\!P^\flat(A^\flat(\mathtt{x}))$ in $\mathtt{x}$ that $\mathtt{x}^\flat$ from (4.5.4) belongs to the interval $[\mathtt{x}^-, \mathtt{x}^+]$ and

$$A^-(\mathtt{x}^-) \subseteq A^\flat(\mathtt{x}^-) \subseteq A^\flat(\mathtt{x}^\flat) \subseteq A^\flat(\mathtt{x}^+) \subseteq A^+(\mathtt{x}^+).$$

This implies the result. □

Now we are prepared to complete the proof. The relations (4.5.11) and (4.5.8) imply

$$I\!P^+\big(A^\flat(\mathtt{x}^\flat)\big) \le I\!P^+\big(A^+(\mathtt{x}^+)\big) = I\!P^-\big(A^-(\mathtt{x}^+)\big).$$

Furthermore, it holds by Corollary 5.3.2 in view of (4.5.5) and (4.5.10)

$$I\!P^-\big(A^-(\mathtt{x}^+)\big) \le I\!P^+\big(A^-(\mathtt{x}^+)\big) + \Delta \le 1 - \exp(-\mathtt{x}) + \Delta.$$

Similarly

$$\begin{aligned} I\!P^-\big(A^\flat(\mathtt{x}^\flat)\big) &\ge I\!P^-\big(A^-(\mathtt{x}^-)\big) = I\!P^+\big(A^+(\mathtt{x}^-)\big) \\ &\ge I\!P^-\big(A^+(\mathtt{x}^-)\big) - \Delta = 1 - \exp(-\mathtt{x}) - \Delta. \end{aligned}$$

This implies (4.5.6) for the sandwiched measure $I\!P$. □

Now we have to show the form of the bound for the operator norm:

$$\|S^{-1/2}\, S^\flat\, S^{-1/2} - \mathbf{1}_p\|_{\mathrm{op}} \le \delta.$$

This can be done by using Theorem 5.1.6 and Proposition 5.1.5.

$$\begin{aligned} \|S^{-1/2}\, S^\flat\, S^{-1/2} - \mathbf{1}_p\|_{\mathrm{op}} \le{}& \|S^{-1/2}\,(S^\flat - S^{\flat'})\, S^{-1/2}\|_{\mathrm{op}} \\ &+ \|S^{-1/2}\, S^{\flat'}\, S^{-1/2} - \mathbf{1}_p\|_{\mathrm{op}}, \end{aligned}$$

where we have defined $S^{\flat'} \stackrel{\mathrm{def}}{=} \Psi\, \mathrm{diag}\big(\boldsymbol{Y}' \cdot \boldsymbol{Y}'\big)\Psi^\top$ with $\boldsymbol{Y}'$ defined in (4.5.2). The first term represents the payment for ignoring the dependency structure of the residuals $\check{\boldsymbol{\varepsilon}}$. The second term can then be represented as a sum of independent matrices. To bound the first term we can apply Proposition 5.1.5 and for the second, we use Theorem 5.1.6. This gives us on a set $\Omega(\mathtt{x})$ of probability at least $1 - 7\exp(-\mathtt{x})$ the bound:

$$\begin{aligned} \delta = 2\sqrt{2}\delta_1^2\sqrt{\mathtt{x}_n} + 4\sqrt{2}\delta_1\mathtt{x}_n + 4\sqrt{2}\|\boldsymbol{B}\|_\infty\delta_1\sqrt{\mathtt{x}_n} + 2\delta_\Psi\sqrt{\mathtt{x} + \log(2p)} \\ +2\delta_\Psi^2(\mathtt{x} + \log(2p)) + \|\boldsymbol{B}\|_\infty^2 + \delta_\Psi^2\|\boldsymbol{B}\|\sqrt{2\mathtt{x}}. \end{aligned}$$

So, we can bound

$$|(1 - \exp(-\mathtt{x})) - I\!P\big(A^\flat(\mathtt{x}^\flat)\big)| \le \sqrt{p}\cdot\delta$$

on $\Omega(\mathtt{x})$. Writing $\Delta_{\boldsymbol{\xi}}(\mathtt{x}) = \sqrt{p}\cdot\delta$, then implies the result of the theorem. This gives the final result.

We remark that only in the last steps have we used the specific random

structure of $A^\flat(\mathtt{x}^\flat)$ via the random matrix $S^\flat$. If we consider $S^\flat$ to be a misspecified deterministic covariance matrix supplied with a bound δ for (4.5.5), one gets a similar result for the case of a misspecified covariance structure.

4.5.2 Proof of Theorem 4.3.2

For a fixed pair $m > m^\circ$ from $\mathcal{M}$, consider $\mathtt{p}^\flat_{m,m^\circ} = I\!\!E^\flat\|\boldsymbol{\xi}^\flat_{m,m^\circ}\|^2$ and $\mathtt{p}_{m,m^\circ} = I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2$. The definitions (4.1.3) and (4.5.1) imply

$$\begin{aligned}\boldsymbol{\xi}^\flat_{m,m^\circ} &= \mathcal{K}_{m,m^\circ}\,\mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat = \mathcal{K}_{m,m^\circ}\Sigma^{1/2}\Sigma^{-1/2}\,\mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{\varepsilon}^\flat \\ &= \mathcal{U}_{m,m^\circ}\,\mathrm{diag}(\check{\boldsymbol{\varepsilon}} + \boldsymbol{B})\boldsymbol{\varepsilon}^\flat,\end{aligned}$$

where $\mathcal{U}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathcal{K}_{m,m^\circ}\Sigma^{1/2}$. It holds for $\mathtt{p}^\flat_{m,m^\circ}$:

$$\mathtt{p}^\flat_{m,m^\circ} = I\!\!E^\flat\big\|\boldsymbol{\xi}^\flat_{m,m^\circ}\big\|^2 = \mathrm{tr}\Big(\mathcal{U}_{m,m^\circ}\,\mathrm{diag}\big\{(\check{\boldsymbol{\varepsilon}} + \boldsymbol{B})\cdot(\check{\boldsymbol{\varepsilon}} + \boldsymbol{B})\big\}\,\mathcal{U}^\top_{m,m^\circ}\Big),$$

while $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\Sigma^{1/2}\boldsymbol{\varepsilon}$ and

$$\mathtt{p}_{m,m^\circ} = I\!\!E\big\|\boldsymbol{\xi}_{m,m^\circ}\big\|^2 = \mathrm{tr}\big(\mathcal{U}_{m,m^\circ}\,\mathcal{U}^\top_{m,m^\circ}\big).$$

As we are interested in the ratio $\mathtt{p}^\flat_{m,m^\circ}/\mathtt{p}_{m,m^\circ}$, one can assume, without loss of generality, that $\|\mathcal{U}_{m,m^\circ}\,\mathcal{U}^\top_{m,m^\circ}\|_{\mathrm{op}} = 1$ and $\mathtt{p}_{m,m^\circ} \geq 1$. Now we look at the following decomposition:

$$\mathcal{B} = \mathcal{U}_{m,m^\circ}\,\mathrm{diag}\left((\check{\boldsymbol{\varepsilon}} + \boldsymbol{B})\cdot(\check{\boldsymbol{\varepsilon}} + \boldsymbol{B}) - \mathbf{1}_n\right)\mathcal{U}^\top_{m,m^\circ}.$$

We can apply Theorem 5.1.7 to show that on a set $\Omega(\mathtt{x})$ with $I\!\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 3\exp(-\mathtt{x})$

$$\left|\frac{\mathtt{p}^\flat_{m,m^\circ}}{\mathtt{p}_{m,m^\circ}} - 1\right| \leq \|\boldsymbol{B}\|^2_\infty + 4\,\mathtt{x}^{1/2}\,\delta_n^2\,\|\boldsymbol{B}\| + 4\mathtt{x}^{1/2}\,\delta_n + 4\,\mathtt{x}\,\delta_n^2 + \delta_2\,.$$

The choice of $\mathtt{x} = \mathtt{x}_{\mathcal{M}} = \mathtt{x} + 2\log\big(\big|\mathcal{M}\big|\big)$ ensures a uniform bound for all pairs $m > m^\circ$ from $\mathcal{M}$.

4.5.3 Proof of Proposition 4.3.3

By Theorem 5.2.1 and a Bonferroni correction, we have for $\mathtt{x}_{\mathcal{M}} = \mathtt{x} + 2\log(|\mathcal{M}|)$.

$$\mathbb{P}^\flat\left(\max_{m>m^\circ, m, m^\circ \in \mathcal{M}} \left\{ \|\boldsymbol{\xi}^\flat_{m,m^\circ}\| - (1+\beta)\sqrt{\mathtt{p}^\flat_{m,m^\circ}} - \sqrt{2\lambda^\flat_{m,m^\circ}\mathtt{x}_{\mathcal{M}}} \right\} \geq 0 \right) \leq \exp(-\mathtt{x}) \quad . \tag{4.5.13}$$

We now want to replace the boostrap-world quantities $\mathtt{p}^\flat_{m,m^\circ}, \lambda^\flat_{m,m^\circ}$ by their real world counterparts. For the effective dimension, we have, due to Theorem 4.3.2, the following bound:

$$|\mathtt{p}_{m,m^\circ} - \mathtt{p}^\flat_{m,m^\circ}| \leq \Delta_{\mathtt{p}} \mathtt{p}_{m,m^\circ}$$

for all $m > m^\circ, m, m^\circ \in \mathcal{M}$. Now we bound, using the definitions of the proof of Theorem 4.3.2:

$$|\lambda_{m,m^\circ} - \lambda^\flat_{m,m^\circ}| \leq \|\mathcal{U}_{m,m^\circ}\left(\mathrm{diag}\big\{(\check{\boldsymbol{\varepsilon}} + \boldsymbol{B}) \cdot (\check{\boldsymbol{\varepsilon}} + \boldsymbol{B})\big\} - \mathbf{1}_n\right)\mathcal{U}^\top_{m,m^\circ}\|_{\mathrm{op}}.$$

As we are only interested in the ratio $\lambda^\flat_{m,m^\circ}/\lambda_{m,m^\circ}$, we assume without loss of generality that $\lambda_{m,m^\circ} = \|\mathcal{U}_{m,m^\circ}\mathcal{U}^\top_{m,m^\circ}\|_{\mathrm{op}} = 1$. Now we can apply Theorem 5.1.6 and Theorem 5.1.5 to get:

$$|\lambda_{m,m^\circ} - \lambda^\flat_{m,m^\circ}| \leq \Delta_{\boldsymbol{\xi}} \lambda_{m,m^\circ}$$

for all $m > m^\circ, m, m^\circ \in \mathcal{M}$. Plugging the bounds into (4.5.13) completes the proof.

4.5.4 Proof of Theorem 4.3.5

We start by showing that

$$\mathbb{P}\left(\widehat{m} > m^*\right) \leq 11\exp(-\mathtt{x}) + \Delta_{\boldsymbol{\xi}}(\mathtt{x}).$$

We write

$$
\begin{aligned}
&I\!P\left(\widehat{m} > m^*\right) \\
&\qquad \leq I\!P\left(\max_{m \in \mathcal{M}^+(m^*)} \left\{ \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*}\| - \mathfrak{z}^{\flat}_{m,m^*} \right\} > 0 \right) \\
&\qquad \leq I\!P\left(\max_{m \in \mathcal{M}^+(m^*)} \left\{ \|\boldsymbol{\xi}_{m,m^*}\| - z^{\flat,+}_{m,m^*}(\mathtt{x}) \right\} > 0 \right) \\
&\qquad + I\!P\left(\max_{m > m^*} \left\{ \|\boldsymbol{b}_{m,m^\circ}\|^2 - \beta^2 \mathtt{p}^{\flat}_{m,m^\circ} \right\} > 0 \right) \\
&\qquad \leq 11 \exp(-\mathtt{x}) + \Delta_{\boldsymbol{\xi}}(\mathtt{x}).
\end{aligned}
$$

The second inequality follows from the triangle inequality and the definitions (3.2.4) and (4.2.3). The last inequality follows by application of Theorem 4.3.1 and Theorem 4.3.2 with the oracle property (4.3.3). Now the rest of the proof follows along the lines of the proof of Theorem 3.4.1. We can again use that $\widehat{m}$ is accepted, which implies by definition

$$
\left\| \widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*} \right\| \leq \mathfrak{z}^{\flat}_{m^*,m} \,.
$$

This yields (4.3.4) and the bound (4.3.5) follows by the triangle inequality.

Chapter 5

Technical results

5.1 Concentration inequalities for norms and traces of a class of random matrices

In this section, we collect and prove a number of deviation bounds for sums of random matrices in different norms. They are mainly used as technical tools to show the validity of replacing unknown deterministic quantities by their bootstrap counterparts.

We start by stating a result from the literature about random matrix bounds, which can be found in [Tropp, 2015]. We are going to adapt the result to our special needs and finally we will present some possibly new bounds for matrix norms and traces of a special class of matrices appearing in our proofs. We recall the matrix norms we are going to use in the following: The operator norm of a square-matrix $A \in \mathbb{R}^{p\times p}$ is defined as

$$\|A\|_{\text{op}} \stackrel{\text{def}}{=} \sqrt{\sup_{\gamma\in\mathbb{R}^p} \frac{\gamma^\top A^\top A\gamma}{\|\gamma\|^2}},$$

which is just the largest (in magnitude) eigenvalue of A. We are also considering the Frobenius norm of the matrix A which is defined as:

$$\|A\|_{\text{Fr}} \stackrel{\text{def}}{=} \sqrt{\text{tr}(A^\top A)}.$$

This can be seen to be the standard Euclidean norm if one considers A as a vector, stripping it of its matrix structure, $\|A\|_{\text{Fr}}^2 = \sum_{i=1}^p \sum_{j=1}^p A_{ij}^2$. The

two norms are related in the following way:

$$\|A\|_{\mathrm{Fr}} \leq p\|A\|_{\mathrm{op}}, \quad \|A\|_{\mathrm{op}} \leq \|A\|_{\mathrm{Fr}}.$$

Master bound

A main ingredient of the proofs will be the following so-called "master bound" (Thm. 3.6.1, [Tropp, 2015]), which gives a concentration result for the algebraically largest eigenvalue $\lambda_{\max}^{+}(\sum_{i=1}^{n} \mathsf{A}_i)$ of a sum random independent Hermitian matrices as a generalization of Chebyshev's inequality to a random matrix framework.

Theorem 5.1.1 (Master bound, Thm. 3.6.1, [Tropp, 2015])**.** *Assume that* $\mathsf{A}_1, \ldots, \mathsf{A}_n \in \mathbb{R}^{p\times p}$ *are independent random Hermitian matrices and* $\boldsymbol{Z} = \sum_{i=1}^{n} \mathsf{A}_i$ *. Then*

$$I\!P\{\lambda_{\max}^{+}(\boldsymbol{Z}) \geq z\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr}\exp\left(\sum_{i=1}^{n} \log I\!E\mathrm{e}^{\theta \mathsf{A}_i}\right).$$

Applying the result to $-\boldsymbol{Z}$ as-well yields a bound for the operator norm $\|\boldsymbol{Z}\|_{\mathrm{op}}$:

$$\begin{aligned} I\!P\{\|\boldsymbol{Z}\|_{\mathrm{op}} \geq z\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr}\exp\left(\sum_{i=1}^{n} \log I\!E\mathrm{e}^{\theta \mathsf{A}_i}\right) \\ + \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr}\exp\left(\sum_{i=1}^{n} \log I\!E\mathrm{e}^{-\theta \mathsf{A}_i}\right). \end{aligned} \tag{5.1.1}$$

Bounds for the operator-norm

We are first going to consider bounds on the operator norm of certain matrices. The next result provides a type of Matrix-Bernstein inequality for the operator norm of a matrix-valued quadratic form.

Theorem 5.1.2. *Consider a matrix* $\mathcal{U} \in I\!R^{p\times n}$ *such that*

$$\mathcal{U}\mathcal{U}^{\top} = \mathbf{1}_p.$$

Let the columns $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n \in I\!R^p$ *of the matrix* $\mathcal{U}$ *satisfy*

$$\|\boldsymbol{\omega}_i\| \leq \delta_n, 1 \leq i \leq n, \tag{5.1.2}$$

for a fixed constant $\delta_n > 0$. *For a random vector* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ *with independent standard normal components, define*

$$\boldsymbol{Z} \stackrel{\mathrm{def}}{=} \mathcal{U} \operatorname{diag}(\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon} - 1) \mathcal{U}^\top = \sum_{i=1}^{n} (\varepsilon_i^2 - 1) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top .$$

Then

$$I\!P\Big(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq 2\delta_n \sqrt{\mathtt{x} + \log(2p)} + 2\delta_n^2 (\mathtt{x} + \log(2p)) \Big) \leq \exp(-\mathtt{x}).$$

Proof. From the Master bound (5.1.1) we get

$$\begin{aligned} I\!P\big(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq z \big) \leq & \inf_{\theta > 0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp \left(\sum_{i=1}^{n} \log I\!E \exp(\theta (\varepsilon_i^2 - 1) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top) \right) \\ & + \inf_{\theta > 0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp \left(\sum_{i=1}^{n} \log I\!E \exp(\theta (-\varepsilon_i^2 + 1) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top) \right). \end{aligned} \tag{5.1.3}$$

We use the following general result:

Lemma 5.1.3. *If* ξ *is a random variable and* Π *is a projector in* $I\!R^p$, *then*

$$\log I\!E \exp(\xi \Pi) = \log\big(I\!E \exp(\xi) \big) \Pi. \tag{5.1.4}$$

Proof. The result (5.1.4) can be easily obtained by applying twice the spectral mapping theorem. □

This result yields, in particular, for any unit vector $\boldsymbol{\omega} \in I\!R^p$

$$\log I\!E \exp\big(\xi \boldsymbol{\omega} \boldsymbol{\omega}^\top \big) = \log\big(I\!E \exp(\xi) \big) \boldsymbol{\omega} \boldsymbol{\omega}^\top .$$

Moreover, for any non-zero vector $\boldsymbol{\omega} \in I\!R^p$, the normalized product $\boldsymbol{\omega}\boldsymbol{\omega}^\top / \|\boldsymbol{\omega}\|^2$ is a rank-one projector, and hence,

$$\log I\!E \exp\big(\xi \boldsymbol{\omega} \boldsymbol{\omega}^\top \big) = \log\big(I\!E \mathrm{e}^{\xi \|\boldsymbol{\omega}\|^2} \big) \frac{\boldsymbol{\omega} \boldsymbol{\omega}^\top}{\|\boldsymbol{\omega}\|^2} .$$

With $\boldsymbol{U}_i \stackrel{\text{def}}{=} \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top / \|\boldsymbol{\omega}_i\|^2$ and $\xi_i = \theta(\varepsilon_i^2 - 1)$ for $1 \le i \le n$, we derive for $\theta \le (2\delta_n^2)^{-1}$:

$$\begin{aligned}
\log I\!\!E \exp\Big(\theta(\varepsilon_i^2 - 1)\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top\Big) &= \log I\!\!E \exp\big(\theta(\varepsilon_i^2 - 1)\|\boldsymbol{\omega}_i\|^2\big)\,\boldsymbol{U}_i \\
&= \log\left(\frac{\exp\bigl(-\|\boldsymbol{\omega}_i\|^2\theta\bigr)}{\sqrt{1 - 2\|\boldsymbol{\omega}_i\|^2\theta}}\right)\boldsymbol{U}_i \\
&= \left(-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\right)\boldsymbol{U}_i
\end{aligned}$$

and

$$\begin{aligned}
\log I\!\!E \exp\Big(\theta(-\varepsilon_i^2 + 1)\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top\Big) &= \log I\!\!E \exp\big(\theta(-\varepsilon_i^2 + 1)\|\boldsymbol{\omega}_i\|^2\big)\,\boldsymbol{U}_i \\
&\le -\|\boldsymbol{\omega}_i\|^2\theta\boldsymbol{U}_i \\
&\le \left(-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\right)\boldsymbol{U}_i
\end{aligned}$$

and therefore by (5.1.3):

$$\begin{aligned}
&I\!\!P\bigl(\|\boldsymbol{Z}\|_{\mathrm{op}} \ge z\bigr) \\
&\quad \le 2 \inf_{\theta > 0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(\sum_{i=1}^n \frac{\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top}{\|\boldsymbol{\omega}_i\|^2}\Big\{-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\Big\}\right). \qquad (5.1.5)
\end{aligned}$$

Denote $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$, where

$$\eta_i = -\theta - \frac{\log(1 - 2\|\boldsymbol{\omega}_i\|^2\theta)}{2\|\boldsymbol{\omega}_i\|^2}.$$

One can show that for any $a < 1$, we have the bound:

$$-a - \log(1 - a) \le \frac{a^2}{2(1 - a)},$$

and therefore (5.1.2) yields for $\theta < (2\delta_n^2)^{-1}$

$$\begin{aligned}
\eta_i &= \frac{1}{2\|\boldsymbol{\omega}_i\|^2}\big\{2\theta\|\boldsymbol{\omega}_i\|^2 - \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\big\} \\
&\le \frac{\bigl(2\theta\|\boldsymbol{\omega}_i\|^2\bigr)^2}{4\|\boldsymbol{\omega}_i\|^2(1 - 2\theta\delta_n^2)} \le \frac{\theta^2\delta_n^2}{1 - 2\theta\delta_n^2}.
\end{aligned}$$

Then by (5.1.5) and $\mathcal{U}\mathcal{U}^\top = \mathbf{1}_p$, using $\mu = 2\theta\delta_n^2$,

$$\begin{aligned}\mathbb{P}\big(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq z\big) &\leq 2 \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\{\mathcal{U} \operatorname{diag}(\boldsymbol{\eta}) \mathcal{U}^\top\} \leq 2 \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\{\|\boldsymbol{\eta}\|_\infty \mathbf{1}_p\} \\ &\leq 2p \inf_{\theta>0} \exp\left\{-\theta z + \frac{\theta^2\delta_n^2}{1-2\theta\delta_n^2}\right\} = 2p \inf_{\mu>0} \exp\left\{-\mu\frac{z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1-\mu}\right\}.\end{aligned}$$

Now we note that one can also show by some algebra, that for $\mathtt{v} > 0$ and $\mathtt{x} > 0$, it holds

$$\inf_{\mu>0}\left\{-\mu\big(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\big) + \frac{\mu^2\mathtt{v}^2}{4(1-\mu)}\right\} \leq -\mathtt{x}.$$

And for $\mathtt{x}_p = \mathtt{x} + \log(2p)$ and $z = 2\delta_n \mathtt{x}_p^{1/2} + 2\delta_n^2 \mathtt{x}_p$ we therefore have:

$$\inf_{\mu>0} \exp\left\{-\mu\frac{z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1-\mu}\right\} = \inf_{\mu>0}\left\{-\mu\big(\delta_n^{-1}\mathtt{x}_p^{1/2} + \mathtt{x}_p\big) + \frac{\mu^2\delta_n^{-2}}{4(1-\mu)}\right\} \leq -\mathtt{x}_p\,.$$

It follows

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_{\mathrm{op}} \geq 2\delta_n\sqrt{\mathtt{x}_p} + 2\delta_n^2\mathtt{x}_p\right) \leq 2p\,\mathrm{e}^{-\mathtt{x}_p} = \exp(-\mathtt{x})$$

as required. □

In the next theorem, we give a deviation for a sum of deterministic matrices weighted by independent Gaussian coefficients. We already specify a setup, which will be suited for our applications. General results in this direction can be found in Chapter 4 of [Tropp, 2015].

Theorem 5.1.4. *Let the vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n$ in $\mathbb{R}^p$ satisfy*

$$\|\boldsymbol{\omega}_i\| \leq \delta_n$$

for a fixed constant δ_n. Let the $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n)$ be a standard normal vector. Then for each vector $\boldsymbol{B} = (b_1, \ldots, b_n)^\top \in \mathbb{R}^n$, the matrix $\boldsymbol{Z}_1$ with

$$\boldsymbol{Z}_1 \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \varepsilon_i b_i \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top$$

satisfies

$$\mathbb{P}\left(\|\boldsymbol{Z}_1\|_{\mathrm{op}} \geq \delta_n^2 \|\boldsymbol{B}\| \sqrt{2\mathtt{x}}\right) \leq 2\exp(-\mathtt{x}).$$

Proof. As the coordinates ε_i of $\boldsymbol{\varepsilon}$ are i.i.d. standard normal and $I\!E \mathrm{e}^{a\varepsilon_i} = \mathrm{e}^{a^2/2}$ for $|a| < 1/2$ and also $\boldsymbol{Z}_1$ has a symmetric distribution: $\boldsymbol{Z}_1 \sim -\boldsymbol{Z}_1$, it follows from the Master inequality and Lemma 5.1.3 that

$$I\!P\bigl(\|\boldsymbol{Z}_1\|_{\mathrm{op}} \geq z\bigr) \leq 2 \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log I\!E \exp(\theta \varepsilon_i b_i \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top) \right)$$
$$\leq 2 \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \frac{\theta^2 b_i^2 \|\boldsymbol{\omega}_i\|^4}{2} \, \frac{\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top}{\|\boldsymbol{\omega}_i\|^2} \right).$$

Moreover, as $\|\boldsymbol{\omega}_i\| \leq \delta_n$ and $\boldsymbol{U}_i = \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top / \|\boldsymbol{\omega}_i\|^2$ is a rank-one projector with $\operatorname{tr} \boldsymbol{U}_i = 1$, it holds

$$\operatorname{tr} \exp\left(\frac{\theta^2}{2} \sum_{i=1}^{n} b_i^2 \|\boldsymbol{\omega}_i\|^4 \boldsymbol{U}_i \right) \leq \exp \operatorname{tr}\left(\frac{\theta^2 \delta_n^4}{2} \sum_{i=1}^{n} b_i^2 \boldsymbol{U}_i \right) = \exp\left(\frac{\theta^2 \delta_n^4 \|\boldsymbol{B}\|^2}{2} \right)$$

by the positive-definiteness of the $\boldsymbol{U}_i$ and the use of Lemma 5.1.3 to justify the exchange of trace and exponential. This implies for $z = \delta_n^2 \|\boldsymbol{B}\| \sqrt{2\mathtt{x}}$

$$I\!P\bigl(\|\boldsymbol{Z}_1\|_{\mathrm{op}} \geq z\bigr) \leq 2 \inf_{\theta>0} \exp\left(-\theta z + \frac{1}{2} \theta^2 \delta_n^4 \|\boldsymbol{B}\|^2 \right) = 2 \exp(-\mathtt{x})$$

and the assertion follows. □

Now we present a bound on a type of random matrix we are considering in the main body of the work: Assume given a vector $\boldsymbol{B}$ in $I\!R^n$, and a matrix $\mathcal{U} \in I\!R^{p \times n}$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n)$. We define

$$\mathcal{B} \stackrel{\mathrm{def}}{=} \mathcal{U} \operatorname{diag}\left((\boldsymbol{\varepsilon} + \boldsymbol{B}) \cdot (\boldsymbol{\varepsilon} + \boldsymbol{B}) - \mathbf{1}_n \right) \mathcal{U}^\top,$$

where $\boldsymbol{u} \cdot \boldsymbol{w} = (u_i * w_i)_{1 \leq i \leq n}$ denotes a coordinate-wise product for two vectors $\boldsymbol{u}, \boldsymbol{w} \in I\!R^n$. For the next theorem, we also assume given another centered normal vector $\breve{\boldsymbol{\varepsilon}}$ on the same space which has a different possibly non-iid covariance matrix, such that $\sup_{1 \leq i \leq n} \operatorname{Var}(\varepsilon_i - \breve{\varepsilon}_i) \leq \delta_n^2$ for some $\delta_n \geq 0$.

Proposition 5.1.5. *Assume that* $\sup_{1 \leq i \leq n} \operatorname{Var}(\varepsilon_i - \breve{\varepsilon}_i) \leq \delta_n^2$ *and* $\|\mathcal{U}^\top \mathcal{U}\|_{\mathrm{op}} \leq 1$. *Then on a random set* $\Omega(\mathtt{x})$ *with* $I\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - 4\exp(-\mathtt{x})$, *it holds:*

$$\|\mathcal{U}\left(\operatorname{diag}(\breve{\boldsymbol{X}} \cdot \breve{\boldsymbol{X}}) - \operatorname{diag}(\boldsymbol{X}' \cdot \boldsymbol{X}') \right) \mathcal{U}^\top\|_{\mathrm{op}} \leq$$
$$4\delta_n^2 \mathtt{x}_n + 4\sqrt{2}\delta_n \mathtt{x}_n + 4\sqrt{2}\|\boldsymbol{B}\|_\infty \delta_n \sqrt{\mathtt{x}_n}\,.$$

Proof. We write

$$\check{\boldsymbol{X}} \stackrel{\text{def}}{=} \check{\boldsymbol{\varepsilon}} + \boldsymbol{B}.$$

and

$$\boldsymbol{X} \stackrel{\text{def}}{=} \boldsymbol{\varepsilon} + \boldsymbol{B}.$$

Then we bound:

$$\begin{aligned}\left\| \mathcal{U} \left(\operatorname{diag}(\check{\boldsymbol{X}} \cdot \check{\boldsymbol{X}}) - \operatorname{diag}(\boldsymbol{X} \cdot \boldsymbol{X}) \right) \mathcal{U}^\top \right\|_{\text{op}} &\leq \left\| \mathcal{U} \left(\operatorname{diag}(\check{\boldsymbol{\varepsilon}} \cdot \check{\boldsymbol{\varepsilon}}) - \operatorname{diag}(\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}) \right) \mathcal{U}^\top \right\|_{\text{op}} \\ &\quad + 2 \left\| \mathcal{U} \left(\operatorname{diag}(\boldsymbol{B} \cdot \check{\boldsymbol{\varepsilon}}) - \operatorname{diag}(\boldsymbol{B} \cdot \boldsymbol{\varepsilon}) \right) \mathcal{U}^\top \right\|_{\text{op}}\end{aligned}$$

We start by considering the first term and note that by the following algebraic identity: $a^2 - b^2 = (a-b)^2 - 2b(b-a), a, b \in I\!R$ we can write:

$$\begin{aligned}\|\mathcal{U}(\operatorname{diag}(\check{\boldsymbol{\varepsilon}} \cdot \check{\boldsymbol{\varepsilon}}) - \operatorname{diag}(\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}))\mathcal{U}^\top\|_{\text{op}} &\leq \|\mathcal{U} \operatorname{diag}((\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}) \cdot (\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}))\mathcal{U}^\top\|_{\text{op}} \\ &\quad + 2\|\mathcal{U} \operatorname{diag}(\boldsymbol{\varepsilon} \cdot (\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}))\mathcal{U}^\top\|_{\text{op}}.\end{aligned}$$

Now $\check{\boldsymbol{\varepsilon}}$ enters into the bound only in the difference $\check{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}$. We write D_+ resp. $-D_-$ for the positive and negative part of $\operatorname{diag}(\boldsymbol{\varepsilon} \cdot (\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}))$. We can write

$$\begin{aligned}\|\mathcal{U} \operatorname{diag}(\boldsymbol{\varepsilon} \cdot (\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}))\mathcal{U}^\top\|_{\text{op}} &\leq \|\mathcal{U} D_+ \mathcal{U}^\top\|_{\text{op}} + \|\mathcal{U} D_- \mathcal{U}^\top\|_{\text{op}} \\ &\leq \|D_+^{1/2} \mathcal{U}^\top \mathcal{U} D_+^{1/2}\|_{\text{op}} + \|D_-^{1/2} \mathcal{U}^\top \mathcal{U} D_-^{1/2}\|_{\text{op}} \\ &\leq (\|D_+\|_{\text{op}} + \|D_-\|_{\text{op}}) \|\mathcal{U}^\top \mathcal{U}\|_{\text{op}} \leq 2\|\boldsymbol{\varepsilon}\|_\infty \|\check{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}\|_\infty,\end{aligned}$$

where we have used the well-known fact that for a rectangular matrix M the sets of non-zero eigenvalues of $M^\top M$ and $M M^\top$ are the same and the assumption that $\|\mathcal{U}\mathcal{U}^\top\|_{\text{op}} \leq 1$. Similarly, we get

$$\begin{aligned}\|\mathcal{U} \operatorname{diag}((\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}) \cdot (\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}))\mathcal{U}^\top\|_{\text{op}} &\leq \|\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}\|_\infty^2, \\ \|\mathcal{U}(\operatorname{diag}(\boldsymbol{B} \cdot \check{\boldsymbol{\varepsilon}}) - \operatorname{diag}(\boldsymbol{B} \cdot \boldsymbol{\varepsilon}))\mathcal{U}^\top\|_{\text{op}} &\leq 2\|\boldsymbol{B}\|_\infty \|\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}\|_\infty.\end{aligned}$$

We now want to give uniform bounds for the random quantities involved. By the assumption on the variance of $\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}$, we then have:

$$I\!P(\|\boldsymbol{\varepsilon} - \check{\boldsymbol{\varepsilon}}\|_\infty \geq \delta_n \sqrt{2\mathtt{x}_n}) \leq 2\exp(-\mathtt{x})$$

with $\mathtt{x}_n = \mathtt{x} + \log(n)$ using a Bonferroni bound and similarly

$$I\!P(\|\boldsymbol{\varepsilon}\|_\infty \geq \sqrt{2\mathtt{x}_n}) \leq 2\exp(-\mathtt{x}).$$

Combining the bounds gives us the statement.

□

We continue to work with the same matrix

$$\mathcal{B} \stackrel{\text{def}}{=} \mathcal{U}\,\mathrm{diag}\left((\boldsymbol{\varepsilon}+\boldsymbol{B})\cdot(\boldsymbol{\varepsilon}+\boldsymbol{B}) - \mathbf{1}_n\right)\mathcal{U}^\top.$$

Now we show a bound on the operator norm of such a matrix in the case of independent errors.

Theorem 5.1.6. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n)$, $\|\mathcal{U}\mathcal{U}^\top\|_{\mathrm{op}} \leq \mathbf{1}_p$ *and the vectors* $\boldsymbol{\omega}_i$ — *the columns of* $\mathcal{U}$ — *satisfy*

$$\max_{1\leq i\leq n} \|\boldsymbol{\omega}_i\| \leq \delta_n.$$

Then on a random set $\Omega(\mathtt{x})$ *with* $I\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 3\exp(-\mathtt{x})$, *it holds*

$$\big\|\mathcal{B}\big\|_{\mathrm{op}} \leq 2\delta_n\sqrt{\mathtt{x}+\log(2p)} + 2\delta_n^2(\mathtt{x}+\log(2p)) + \|\boldsymbol{B}\|_\infty^2 + \delta_n^2\|\boldsymbol{B}\|\sqrt{2\mathtt{x}}.$$

Proof. We use the representation

$$\begin{aligned}\mathcal{B} &= \mathcal{U}\,\mathrm{diag}\left((\boldsymbol{\varepsilon}+\boldsymbol{B})\cdot(\boldsymbol{\varepsilon}+\boldsymbol{B}) - \mathbf{1}_n\right)\mathcal{U}^\top \\ &= \underbrace{\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon} - \mathbf{1}_n\right)\mathcal{U}^\top}_{\mathcal{B}_1} \\ &\quad + \underbrace{\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{B}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_2} + \underbrace{2\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_3}.\end{aligned}$$

By the triangle inequality:

$$\|\mathcal{B}\|_{\mathrm{op}} \leq \|\mathcal{B}_1\|_{\mathrm{op}} + \|\mathcal{B}_2\|_{\mathrm{op}} + \|\mathcal{B}_3\|_{\mathrm{op}}.$$

We proceed by bounding each $\mathcal{B}_m$ for $m = 1, 2, 3$ separately, starting with $\mathcal{B}_1$. The operator norm $\|\mathcal{B}_1\|_{\mathrm{op}}$ can be bounded by Theorem 5.1.2:

$$I\!P\Big(\|\mathcal{B}_1\|_{\mathrm{op}} \geq 2\delta_n\sqrt{\mathtt{x}+\log(2p)} + 2\delta_n^2(\mathtt{x}+\log(2p))\Big) \leq \exp(-\mathtt{x}).$$

The second term $\|\mathcal{B}_2\|_{\mathrm{op}}$ is bounded deterministically by

$$\|\mathcal{B}_2\|_{\mathrm{op}} = \Big\| \sum_{i=1}^{n} \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top b_i^2 \Big\|_{\mathrm{op}} \leq \|\boldsymbol{B}\|_\infty^2 \|\mathcal{U}\mathcal{U}^\top\|_{\mathrm{op}} \leq \|\boldsymbol{B}\|_\infty^2,$$

using $\|\mathcal{U}\mathcal{U}^\top\|_{\mathrm{op}} \leq 1$. The third term $\|\mathcal{B}_3\|_{\mathrm{op}}$ is bounded by applying Theorem 5.1.4:

$$I\!P\Big(\|\mathcal{B}_3\|_{\mathrm{op}} \geq \delta_n^2 \|\boldsymbol{B}\| \sqrt{2\mathtt{x}} \Big) \leq 2\exp(-\mathtt{x}).$$

Combining the bounds gives us the stated result. □

Bounds for the trace

Now we want to show a concentration result for $\mathrm{tr}(\mathcal{B})$ without the assumption that the coefficients of $\boldsymbol{\varepsilon}$ are independent.

Theorem 5.1.7. *We write*

$$c_1 \stackrel{\mathrm{def}}{=} \|\operatorname{Var}(\boldsymbol{\varepsilon}) - \mathbf{1}_n\|_{\mathrm{op}},$$
$$\delta_2 \stackrel{\mathrm{def}}{=} \max_{1 \leq i \leq n} |I\!E\varepsilon_i^2 - 1|.$$

Suppose that a Gaussian vector $\boldsymbol{\varepsilon} \in I\!R^n$ satisfies

$$c_1 \leq 1.$$

Let also $\mathcal{U}\mathcal{U}^\top \leq \mathbf{1}_p$ and the vectors $\boldsymbol{\omega}_i$ - columns of $\mathcal{U}$ - satisfy for some $\mathtt{q} > 0$:

$$\begin{aligned} \mathrm{tr}(\mathcal{U}\mathcal{U}^\top) = \sum_{i=1}^{n} \|\boldsymbol{\omega}_i\|^2 &\leq \mathtt{q}, \\ \max_{1 \leq i \leq n} \|\boldsymbol{\omega}_i\|^2 &\leq \delta_n^2. \end{aligned} \tag{5.1.6}$$

Then on a random set $\Omega(\mathtt{x})$ with $I\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 3\exp(-\mathtt{x})$, it holds

$$\big|\mathrm{tr}(\mathcal{B})\big| \leq 4\mathtt{x}^{1/2}\sqrt{\mathtt{q}}\,\delta_n + 2\,\mathtt{x}\,\delta_n^2 + \mathtt{q}\,\|\boldsymbol{B}\|_\infty^2 + 4\,\mathtt{x}^{1/2}\,\delta_n^2\,\|\boldsymbol{B}\| + \delta_2\,\mathtt{q}.$$

Proof. We now use the representation

$$\begin{aligned}\mathcal{B} &= \mathcal{U}\,\mathrm{diag}\left((\boldsymbol{\varepsilon}+\boldsymbol{B})\cdot(\boldsymbol{\varepsilon}+\boldsymbol{B})\right)\mathcal{U}^\top - \mathbf{1}_p \\ &= \underbrace{\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon} - I\!\!E(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon})\right)\mathcal{U}^\top}_{\mathcal{B}_1} + \underbrace{\mathcal{U}\,\mathrm{diag}\left(I\!\!E(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}) - \mathbf{1}\right)\mathcal{U}^\top}_{\mathcal{B}_2} \\ &\quad + \underbrace{\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{B}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_3} + \underbrace{2\mathcal{U}\,\mathrm{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_4}.\end{aligned}$$

Again, as in the bound on the operator norm, we start by using the triangle inequality:

$$\bigl|\mathrm{tr}(\mathcal{B})\bigr| \le \bigl|\mathrm{tr}(\mathcal{B}_1)\bigr| + \bigl|\mathrm{tr}(\mathcal{B}_2)\bigr| + \bigl|\mathrm{tr}(\mathcal{B}_3)\bigr| + \bigl|\mathrm{tr}(\mathcal{B}_4)\bigr|.$$

Bound for $\mathcal{B}_1$: This part is most involved because $\mathcal{B}_1$ is a matrix valued quadratic form of $\boldsymbol{\varepsilon}$. By the conditions of the theorem,

$$\|\,\mathrm{Var}(\boldsymbol{\varepsilon}) - \mathbf{1}_n\|_{\mathrm{op}} \le c_1 \le 1. \tag{5.1.7}$$

Therefore, $\|\,\mathrm{Var}(\boldsymbol{\varepsilon})\|_{\mathrm{op}} \le 1 + c_1 \le 2$. By definition, it holds for the columns $\boldsymbol{\omega}_i \in I\!R^q$ of $\mathcal{U}$

$$\mathrm{tr}(\mathcal{B}_1) = \sum_{i=1}^n \mathrm{tr}\bigl(\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top\bigr)(\varepsilon_i^2 - I\!\!E\varepsilon_i^2) = \sum_{i=1}^n \|\boldsymbol{\omega}_i\|^2(\varepsilon_i^2 - I\!\!E\varepsilon_i^2),$$

and by Corollary 5.2.4, it holds on a random set $\Omega(\mathtt{x})$ with $I\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - 2\exp(-\mathtt{x})$

$$\bigl|\mathrm{tr}(\mathcal{B}_1)\bigr| \le 4\mathtt{x}^{1/2}\sqrt{\sum_{i=1}^n \|\boldsymbol{\omega}_i\|^4} + 2\mathtt{x}\max_{1\le i\le n}\|\boldsymbol{\omega}_i\|^2.$$

This implies in view of $\|\boldsymbol{\omega}_i\| \le \delta_n$ and (5.1.6)

$$\sqrt{\sum_{i=1}^n \|\boldsymbol{\omega}_i\|^4} \le \max_{1\le i\le n}\|\boldsymbol{\omega}_i\|\sqrt{\sum_{i=1}^n \|\boldsymbol{\omega}_i\|^2} \le \sqrt{\mathtt{q}}\,\delta_n.$$

Therefore, it holds on $\Omega(\mathtt{x})$

$$\bigl|\mathrm{tr}(\mathcal{B}_1)\bigr| \le 4\mathtt{x}^{1/2}\sqrt{\mathtt{q}}\,\delta_n + 2\,\mathtt{x}\,\delta_n^2.$$

Bound for $\mathcal{B}_2$: By direct calculation, it holds, in view of (5.1.6) and $|\mathbb{E}\varepsilon_i^2 - 1| \le \delta_2$, that

$$\bigl|\mathrm{tr}(\mathcal{B}_2)\bigr| \le \mathrm{tr}\biggl(\sum_{i=1}^n \boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top \bigl|\mathbb{E}\varepsilon_i^2 - 1\bigr|\biggr) \le \delta_2\,\mathrm{tr}(\mathcal{U}\mathcal{U}^\top) \le \delta_2\,\mathtt{q}.$$

Bound for $\mathcal{B}_3$: The bias term $\mathcal{B}_3$ can be estimated in a similar way:

$$|\,\mathrm{tr}(\mathcal{B}_3)| = \mathrm{tr}\biggl(\sum_{i=1}^n \boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top b_i^2\biggr) \le \|\boldsymbol{B}\|_\infty^2\,\mathrm{tr}\bigl(\mathcal{U}\mathcal{U}^\top\bigr) \le \mathtt{q}\,\|\boldsymbol{B}\|_\infty^2.$$

Bound for $\mathcal{B}_4$: We have

$$\mathrm{tr}(\mathcal{B}_4) = 2\,\mathrm{tr}\biggl(\sum_{i=1}^n \boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top \varepsilon_i b_i\biggr) = 2\sum_{i=1}^n \|\boldsymbol{\omega}_i\|^2 \varepsilon_i b_i = 2\boldsymbol{u}^\top\boldsymbol{\varepsilon},$$

where $\boldsymbol{u}$ is the vector in $\mathbb{R}^n$ with the entries $u_i = \|\boldsymbol{\omega}_i\|^2 b_i, 1 \le i \le n$. As $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbb{V}$ with $\|\mathbb{V}\|_{\mathrm{op}} \le 2$, $\boldsymbol{u}^\top\boldsymbol{\varepsilon}$ is a Gaussian zero mean random variable whose variance satisfies

$$\mathrm{Var}(\boldsymbol{u}^\top\boldsymbol{\varepsilon}) \le \boldsymbol{u}^\top\mathbb{V}\boldsymbol{u} \le 2\|\boldsymbol{u}\|^2 = 2\sum_{i=1}^n \|\boldsymbol{\omega}_i\|^4 b_i^2 \le 2\delta_n^4\|\boldsymbol{B}\|^2.$$

Here we have used (5.1.7) and $\|\boldsymbol{\omega}_i\| \le \delta_n$. Therefore, on a random set $\Omega(\mathtt{x})$ with $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - 2\exp(-\mathtt{x})$,

$$|\,\mathrm{tr}(\mathcal{B}_4)| \le 2\sqrt{2}\,\delta_n^2\,\|\boldsymbol{B}\|\,z_1(\mathtt{x}) \le 4\,\mathtt{x}^{1/2}\delta_n^2\,\|\boldsymbol{B}\|,$$

where $z_1(\mathtt{x}) \le \sqrt{2\mathtt{x}}$ is given by $\mathbb{P}(|\xi| > z_1(\mathtt{x})) = \exp(-\mathtt{x})$ for a one-dimensional standard normal ξ. □

Finally, we are also going to present bounds on the Frobenius norm for matrices of type $\mathcal{B}$.

Bounds for the Frobenius norm

Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbb{V})$ be a Gaussian zero mean vector with the covariance matrix $\mathbb{V} \in \mathbb{R}^{n\times n}$ such that $\|\mathbb{V}\|_{\mathrm{op}} = \lambda_{\max}(\mathbb{V}) \le \lambda^*$. Further, let $\mathcal{U} \in \mathbb{R}^{p\times n}$ be a

matrix with columns $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n \in I\!R^p$ such that

$$\begin{aligned} \operatorname{tr}(\mathcal{U}\mathcal{U}^\top) = \sum_{i=1}^n \|\boldsymbol{\omega}_i\|^2 \leq \mathtt{q}, \\ \max_{1 \leq i \leq n} \|\boldsymbol{\omega}_i\|^2 \leq \delta_n. \end{aligned} \tag{5.1.8}$$

A typical situation we have in mind is when

$$\sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top = \mathbf{1}_p.$$

Then (5.1.8) is satisfied with $\mathtt{q} = p$. Moreover, in this case it also holds that $\operatorname{tr}\big((\mathcal{U}\mathcal{U}^\top)^2\big) = \operatorname{tr}(\mathcal{U}\mathcal{U}^\top) = p$.

We aim at establishing a bound on the squared Frobenius norm $\operatorname{tr}(\mathcal{B}^2)$ for the $p \times p$ random symmetric matrix $\mathcal{B}$ of the form

$$\mathcal{B} \stackrel{\text{def}}{=} \sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top (\varepsilon_i^2 - I\!E\varepsilon_i^2). \tag{5.1.9}$$

Theorem 5.1.8. *Let the vectors $\boldsymbol{\omega}_i \in I\!R^p$ satisfy* (5.1.8). *Let also $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbb{V})$ be a zero mean Gaussian vector with $\|\mathbb{V}\|_{\mathrm{op}} \leq \lambda^*$. Then the random matrix $\mathcal{B}$ from* (5.1.9) *satisfies*

$$I\!P\Big(\operatorname{tr}(\mathcal{B}^2) > 4\lambda^* \, \delta_n^2 \, \mathtt{q} \, (\mathtt{x}_n^{1/2} + \delta^* \, \mathtt{x}_n)\Big) \leq \exp(-\mathtt{x}), \tag{5.1.10}$$

where $\delta^* \leq 1$.

Proof. Denote $\zeta_i = \varepsilon_i^2 - I\!E\varepsilon_i^2$ and $c_{ij} = \boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j$ for $1 \leq i, j \leq n$. Then

$$\operatorname{tr}(\mathcal{B}^2) = \sum_{i=1}^n \sum_{j=1}^n \zeta_i \zeta_j \operatorname{tr}\big(\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j \boldsymbol{\omega}_j^\top\big) = \sum_{i=1}^n \sum_{j=1}^n c_{ij}^2 \zeta_i \zeta_j$$

by cyclicity of the trace. The matrix $\boldsymbol{C} = (c_{ij}^2)_{1 \leq i,j \leq n} \in I\!R^{n \times n}$ is obviously symmetric positive-definite. Therefore, one can represent it in the form $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^\top$ for a diagonal matrix $\boldsymbol{M} = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ and an orthonormal $n \times n$ matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ whose columns $\boldsymbol{u}_k$ are orthonormal vectors in $I\!R^n$. Therefore, for the vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^\top$,

$$\operatorname{tr}(\mathcal{B}^2) = \boldsymbol{\zeta}^\top \boldsymbol{C} \boldsymbol{\zeta} = \boldsymbol{\zeta}^\top \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^\top \boldsymbol{\zeta} = \sum_{k=1}^n \mu_k |\boldsymbol{u}_k^\top \boldsymbol{\zeta}|^2. \tag{5.1.11}$$

Further, one can bound each $\boldsymbol{u}_k^\top \boldsymbol{\zeta}$ by the result of Corollary 5.2.4: for any $\mathtt{x}_n > 0$

$$I\!P\Big(|\boldsymbol{u}_k^\top \boldsymbol{\zeta}| > 4\lambda^* \big(\mathtt{x}_n^{1/2} + \|\boldsymbol{u}_k\|_\infty \, \mathtt{x}_n \big) \Big) \leq \mathrm{e}^{-\mathtt{x}_n}.$$

The choice $\mathtt{x}_n = \mathtt{x} + \log(n)$ and (5.1.11) together imply

$$I\!P\bigg(\mathrm{tr}(\mathcal{B}^2) > \sum_{k=1}^n 4\lambda^* \mu_k \big(\mathtt{x}_n^{1/2} + \|\boldsymbol{u}_k\|_\infty \, \mathtt{x}_n \big) \bigg) \leq \exp(-\mathtt{x}).$$

Also by construction and (5.1.8)

$$\sum_{k=1}^n \mu_k = \mathrm{tr}(\boldsymbol{C}) = \sum_{i=1}^n c_{ii}^2 = \sum_{i=1}^n \|\boldsymbol{\omega}_i\|^4 \leq \delta_n^2 \, \mathrm{tr}\bigg(\sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top \bigg) = \delta_n^2 \mathtt{q}.$$

The result (5.1.10) uses a very rough bound $\|\boldsymbol{u}_k\|_\infty \leq \delta^*$ for some constant $\delta^* \leq 1$. In typical situations, one can refine it to $\|\boldsymbol{u}_k\|_\infty \leq \mathtt{C}\delta_n$. □

Now we return to the setting as before: assume given a vector $\boldsymbol{B}$ in $I\!R^n$, and a matrix $\mathcal{U} \in I\!R^{p \times n}$, and assume a mean zero normal vector $\boldsymbol{\varepsilon}$ not necessarily i.i.d., and we write

$$\mathcal{B} \stackrel{\mathrm{def}}{=} \mathcal{U} \, \mathrm{diag}\left((\boldsymbol{\varepsilon} + \boldsymbol{B}) \cdot (\boldsymbol{\varepsilon} + \boldsymbol{B}) - \mathbf{1}_n \right) \mathcal{U}^\top$$

and bound the value of $\|\mathcal{B}\|_{\mathrm{Fr}} = \sqrt{\mathrm{tr}(\mathcal{B}^2)}$. While we will also use the bound of the Frobenius norm via the operator norm, this can be seen as an interesting result in itself giving another independent bound, which could be useful in future research. For example, in the case of a misspecified covariance matrix, one could use arguments for the total variation distance, which do not rely on the sandwiching arguments as in the proof of Theorem 4.3.1 and we could directly use the bound proposed in the following theorem.

Theorem 5.1.9. *Suppose that the Gaussian vector $\boldsymbol{\varepsilon}$ satisfies*

$$c_1 \stackrel{\mathrm{def}}{=} \| \mathrm{Var}(\boldsymbol{\varepsilon}) - \mathbf{1}_n \|_{\mathrm{op}},$$
$$\delta_2^2 \stackrel{\mathrm{def}}{=} \max_{1 \leq i \leq n} |I\!E \varepsilon_i^2 - 1|.$$

Let also $\mathcal{U}\mathcal{U}^\top \le \mathbf{1}_p$ *and the vectors* $\boldsymbol{\omega}_i$ *— columns of* $\mathcal{U}$ *— satisfy for some* $\mathtt{q}_2 \le \mathtt{q}$

$$\begin{aligned} \operatorname{tr}\left((\mathcal{U}\mathcal{U}^\top)^2\right) = \sum_{i,j=1}^{n} \left|\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j\right|^2 &\le \mathtt{q}_2, \\ \max_{1\le i\le n} \|\boldsymbol{\omega}_i\| &\le \delta_n. \end{aligned} \tag{5.1.12}$$

Then on a random set $\Omega(\mathtt{x})$ *with* $I\!P\big(\Omega(\mathtt{x})\big) \ge 1 - 3\exp(-\mathtt{x})$ *, it holds*

$$\|\mathcal{B}\|_{\mathrm{Fr}} = \sqrt{\operatorname{tr}(\mathcal{B}^2)} \le \sqrt{8\,\delta_n^2\,\mathtt{q}\,\mathtt{x}_n} + \sqrt{\delta_2^4\,\mathtt{q}_2} + \sqrt{\|\boldsymbol{B}\|_\infty^4\,\mathtt{q}_2} + 4\delta_n^2\|\boldsymbol{B}\|\big(1+\sqrt{\mathtt{x}}\big).$$

Proof. We use the representation

$$\begin{aligned} \mathcal{B} &= \mathcal{U}\operatorname{diag}\left((\boldsymbol{\varepsilon}+\boldsymbol{B})\cdot(\boldsymbol{\varepsilon}+\boldsymbol{B}) - \mathbf{1}_n\right)\mathcal{U}^\top \\ &= \underbrace{\mathcal{U}\operatorname{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon} - I\!E(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon})\right)\mathcal{U}^\top}_{\mathcal{B}_1} + \underbrace{\mathcal{U}\operatorname{diag}\left(I\!E(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}) - \mathbf{1}\right)\mathcal{U}^\top}_{\mathcal{B}_2} \\ &\quad + \underbrace{\mathcal{U}\operatorname{diag}\left(\boldsymbol{B}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_3} + \underbrace{2\mathcal{U}\operatorname{diag}\left(\boldsymbol{\varepsilon}\cdot\boldsymbol{B}\right)\mathcal{U}^\top}_{\mathcal{B}_4}. \end{aligned}$$

Again with $\|\mathcal{B}\|_{\mathrm{Fr}} \stackrel{\mathrm{def}}{=} \sqrt{\operatorname{tr}(\mathcal{B}^2)}$

$$\|\mathcal{B}\|_{\mathrm{Fr}} \le \|\mathcal{B}_1\|_{\mathrm{Fr}} + \|\mathcal{B}_2\|_{\mathrm{Fr}} + \|\mathcal{B}_3\|_{\mathrm{Fr}} + \|\mathcal{B}_4\|_{\mathrm{Fr}}. \tag{5.1.13}$$

Bound for $\mathcal{B}_1$**:** By the conditions of the theorem,

$$\|\operatorname{Var}(\boldsymbol{\varepsilon}) - \mathbf{1}_n\|_{\mathrm{op}} \le c_1 \le 1.$$

Therefore, $\|\operatorname{Var}(\boldsymbol{\varepsilon})\|_{\mathrm{op}} \le 1 + c_1 \le 2$. By Theorem 5.1.8, for $\mathtt{x}_n = \mathtt{x} + \log(n)$ and $\delta^* \le 1$, it holds on a random set $\Omega(\mathtt{x})$ with $I\!P\big(\Omega(\mathtt{x})\big) \ge 1 - 2\exp(-\mathtt{x})$

$$\operatorname{tr}(\mathcal{B}_1^2) \le 4(1+c_1)\,\delta_n^2\,\mathtt{q}\,(\mathtt{x}_n^{1/2} + \delta^*\,\mathtt{x}_n).$$

Here $\delta^* \le 1$, usually $\delta^* \ll 1$, and $c_1 \le 1$, so we simplify the bound to

$$\operatorname{tr}(\mathcal{B}_1^2) \le 4\,\delta_n^2\,\mathtt{q}\,\mathtt{x}_n.$$

Bound for $\mathcal{B}_2$: Again it holds by direct calculus using (5.1.12) and $|I\!E\varepsilon_i^2 - 1| \le \delta_2^2$:

$$\operatorname{tr}(\mathcal{B}_2^2) \le \operatorname{tr}\biggl(\Bigl(\sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top \bigl|I\!E\varepsilon_i^2 - 1\bigr|\Bigr)^2\biggr) \le \delta_2^4 \operatorname{tr}((\mathcal{U}\mathcal{U}^\top)^2) \le \delta_2^4\, \mathtt{q}_2.$$

Bound for $\mathcal{B}_3$: The bias term $\mathcal{B}_3$ can be estimated in a similar way:

$$\operatorname{tr}(\mathcal{B}_3^2) = \operatorname{tr}\biggl(\Bigl(\sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top b_i^2\Bigr)^2\biggr) \le \|\boldsymbol{B}\|_\infty^4 \operatorname{tr}\Bigl(\bigl(\mathcal{U}\mathcal{U}^\top\bigr)^2\Bigr) \le \mathtt{q}_2\, \|\boldsymbol{B}\|_\infty^4. \tag{5.1.14}$$

Another way of bounding the value $\operatorname{tr}(\mathcal{B}_3^2)$ is based on the fact $\|\mathcal{B}_3\|_{\mathrm{op}} \le \delta_n^2 \|\boldsymbol{B}\|^2$ and the relation of the Frobenius norm and the operator norm,

$$\operatorname{tr}(\mathcal{B}_3^2) \le \bigl(\delta_n^2 \|\boldsymbol{B}\|^2\bigr)^2 p = \delta_n^4 \|\boldsymbol{B}\|^4 p.$$

Note, however, that the bound (5.1.14) is typically more accurate: the value δ_n^2 is of order $\mathtt{q}/n$ and $\|\boldsymbol{B}\|^2 \asymp n\|\boldsymbol{B}\|_\infty^2$, so that $\delta_n^4 \|\boldsymbol{B}\|^4 p \gg \mathtt{q}_2 \|\boldsymbol{B}\|_\infty^4$ for p large.

Bound for $\mathcal{B}_4$: It remains to bound $\operatorname{tr}(\mathcal{B}_4^2)$. Because of cross-dependence of the ε_i's, we cannot directly apply the result of Theorem 5.1.4. Instead we use the following representation:

$$\operatorname{tr}(\mathcal{B}_4^2) = 4 \sum_{i,j=1}^n (\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j)^2 b_i b_j \varepsilon_i \varepsilon_j\,.$$

Denote by $\boldsymbol{C}_1$ the $n \times n$ matrix with the entries $(\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j)^2 b_i b_j$ for $i, j = 1, \ldots, n$. The use of $\boldsymbol{\varepsilon} = \mathbb{V}^{1/2}\boldsymbol{\xi}$ with $\mathbb{V} = \operatorname{Var}(\boldsymbol{\varepsilon})$ and a standard normal $\boldsymbol{\xi} \in I\!R^n$ yields

$$\operatorname{tr}(\mathcal{B}_4^2) = 4\boldsymbol{\varepsilon}^\top \boldsymbol{C}_1 \boldsymbol{\varepsilon} = 4\boldsymbol{\varepsilon}^\top \mathbb{V}^{1/2} \boldsymbol{C}_1 \mathbb{V}^{1/2} \boldsymbol{\varepsilon} = 4\boldsymbol{\varepsilon}^\top \boldsymbol{C}_2 \boldsymbol{\varepsilon},$$

where $\boldsymbol{C}_2 = \mathbb{V}^{1/2} \boldsymbol{C}_1 \mathbb{V}^{1/2}$. Now the bound of Theorem 5.2.1 on Gaussian quadratic forms can be applied. It holds

$$\mathtt{p}(\boldsymbol{C}_2) = \operatorname{tr}(\boldsymbol{C}_2) \le 2 \operatorname{tr}(\boldsymbol{C}_1) = 2 \sum_{i=1}^n \|\boldsymbol{\omega}_i\|^4 b_i^2 \le 2\delta_n^4 \|\boldsymbol{B}\|^2.$$

Similarly for any unit vector $\boldsymbol{u} \in I\!R^n$, it holds by $|\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j| \leq \delta_n^2$

$$\boldsymbol{u}^\top \boldsymbol{C}_1 \boldsymbol{u} = \sum_{i,j=1}^n u_i u_j b_i b_j (\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j)^2 \leq \delta_n^4 \Big(\sum_{i=1}^n u_i b_i \Big)^2 \leq \delta_n^4 \, \|\boldsymbol{u}\|^2 \, \|\boldsymbol{B}\|^2 = \delta_n^4 \, \|\boldsymbol{B}\|^2$$

yielding $\lambda_{\max}(\boldsymbol{C}_1) \leq \delta_n^4 \, \|\boldsymbol{B}\|^2$ and

$$\lambda(\boldsymbol{C}_2) \stackrel{\text{def}}{=} \lambda_{\max}(\boldsymbol{C}_2) \leq 2\delta_n^4 \, \|\boldsymbol{B}\|^2.$$

Theorem 5.2.1 implies on a random set of probability at least $1 - \exp(-\mathtt{x})$

$$\begin{aligned} \sqrt{\operatorname{tr}(\mathcal{B}_4^2)} &= 2\sqrt{\boldsymbol{\varepsilon}^\top \boldsymbol{C}_2 \boldsymbol{\varepsilon}} \\ &\leq 2\sqrt{\mathtt{p}(\boldsymbol{C}_2)} + 2\sqrt{2\lambda(\boldsymbol{C}_2)\mathtt{x}} \\ &= 2\sqrt{2}\, \delta_n^2 \, \|\boldsymbol{B}\| \, \big(1 + \sqrt{2\mathtt{x}}\big) \\ &\leq 4\, \delta_n^2 \, \|\boldsymbol{B}\| (1 + \sqrt{\mathtt{x}}). \end{aligned}$$

Putting all the bounds together yields the statements of the theorem by (5.1.13). □

5.2 Deviation bounds for Gaussian quadratic forms

Here we restate for the sake of self-containedness some results on Gaussian quadratic forms which we use extensively in this work. Similar results and extension can be easily found in the literature [Spokoiny and Zhilova, 2013], [Hsu et al., 2012]. In the following we will present the main deviation bound for quadratic forms and then derive several corollaries from it, which are well-suited to our applications. The next theorem describes the concentration properties of a quadratic form $\boldsymbol{\varepsilon}^\top B \boldsymbol{\varepsilon}$ for a standard normal vector $\boldsymbol{\varepsilon}$ and a symmetric matrix B around its mean $\operatorname{tr}(B)$. We cite [Laurent and Massart, 2000] and specifically Lemma 1 therein as a reference. For the definition of the operator norm and the Frobenius norm, we refer the reader to the notation part of this thesis.

Theorem 5.2.1. *Let $\boldsymbol{\varepsilon}$ be a standard normal Gaussian vector and B be a symmetric non-negative definite matrix. Then with $\mathtt{p} \stackrel{\text{def}}{=} \operatorname{tr}(B)$, $\mathrm{v}^2 \stackrel{\text{def}}{=}$*

$\mathrm{tr}(B^2) = \|B\|_{\mathrm{Fr}}^2$, *and* $\lambda \stackrel{\mathrm{def}}{=} \|B\|_{\mathrm{op}}$, *it holds for each* $\mathtt{x} \geq 0$ *that*

$$I\!P\bigl(\varepsilon^\top B\varepsilon > \mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}\bigr) \leq \exp(-\mathtt{x}).$$

The above bound implies

$$I\!P\bigl(\|B^{1/2}\varepsilon\| > \mathtt{p}^{1/2} + (2\lambda\mathtt{x})^{1/2}\bigr) \leq \exp(-\mathtt{x}). \tag{5.2.1}$$

Additionally, we have:

$$I\!P\bigl(\varepsilon^\top B\varepsilon < \mathtt{p} - 2\mathtt{v}\mathtt{x}^{1/2}\bigr) \leq \exp(-\mathtt{x}).$$

Proof. We normalize by λ to reduce the proof to the case with $\lambda = 1$. An orthogonal transformation and the properties of the multivariate normal distribution let us write the Gaussian quadratic form $\|\varepsilon\|^2$ as a sum of independent χ^2-variables with one degree of freedom:

$$\varepsilon^\top B\varepsilon = \sum_{i=1}^{p} \lambda_i \nu_i^2$$

where the $\nu_i, 1 \leq i \leq p,$ are independent standard normal. Here the $\lambda_i \in [0,1], 1 \leq i \leq p,$ denote the eigenvalues of B, and $\mathtt{p} = \sum_{i=1}^{p} \lambda_i$, $\mathtt{v}^2 = \sum_{i=1}^{p} \lambda_i^2$. The rest of the proof works as stated in Lemma 1, [Laurent and Massart, 2000].

To show the second assertion of the theorem, we note that $\mathrm{tr}(B^2) \leq \|B\|_{\mathrm{op}} \mathrm{tr}(B) = \lambda\,\mathtt{p}$ following from the fact that B is positive-semidefinite. The proof of the lower bound can again be obtained from Lemma 1, [Laurent and Massart, 2000].

□

We can combine the lower and upper bound to get a bound for a symmetric quadratic form $\varepsilon^\top B\varepsilon$, which is not necessarily positive semidefinite.

Corollary 5.2.2. *Let* ε *be standard normal in* $I\!R^p$ *and let* B *be symmetric. Then with* $\mathtt{p} = \mathrm{tr}(B)$, $\mathtt{v}^2 = \mathrm{tr}(B^2)$, *and* $\lambda = \lambda_{\max}(B)$, *it holds for each* $\mathtt{x} \geq 0$

$$I\!P\bigl(\bigl|\varepsilon^\top B\varepsilon - \mathtt{p}\bigr| > 4\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}\bigr) \leq 2\exp(-\mathtt{x}).$$

Proof. As B is symmetric, we can diagonalize and replace $\boldsymbol{\varepsilon}$ by its rotated counterpart $\boldsymbol{\varepsilon}' \sim \mathcal{N}(0, \mathbf{1}_p)$.

$$\boldsymbol{\varepsilon}^\top B \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'^\top D \boldsymbol{\varepsilon}'.$$

with $D \in \mathbb{R}^{p \times p}$ diagonal. Now we split D into its positive semidefinite part D_+ and its negative semidefinite part $-D_-$.

$$\boldsymbol{\varepsilon}'^\top D \boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon}'^\top D_+ \boldsymbol{\varepsilon}' - \boldsymbol{\varepsilon}'^\top D_- \boldsymbol{\varepsilon}'.$$

As D_+ and D_- are positive semidefinite symmetric, we can apply Thm. 5.2.1, which gives us

$$\begin{aligned} &I\!P\Big(\boldsymbol{\varepsilon}'^\top D_+ \boldsymbol{\varepsilon}' - \boldsymbol{\varepsilon}'^\top D_- \boldsymbol{\varepsilon}' \geq \operatorname{tr}(D_+) - \operatorname{tr}(D_-) \\ &\quad + 2\sqrt{\operatorname{tr}(D_+^2)x} + 2\sqrt{\operatorname{tr}(D_-^2)x} + 2\|D_+\|_\infty x\Big) \leq 2\exp(-x), \end{aligned}$$

$$\begin{aligned} &I\!P\Big(\boldsymbol{\varepsilon}'^\top D_+ \boldsymbol{\varepsilon}' - \boldsymbol{\varepsilon}'^\top D_- \boldsymbol{\varepsilon}' \leq \operatorname{tr}(D_+) - \operatorname{tr}(D_-) \\ &\quad + 2\sqrt{\operatorname{tr}(D_+^2)x} + 2\sqrt{\operatorname{tr}(D_-^2)x} + 2\|D_-\|_\infty x\Big) \leq 2\exp(-x). \end{aligned}$$

We now note $\operatorname{tr}(B) = \operatorname{tr}(D_+) - \operatorname{tr}(D_-)$, $\sqrt{\operatorname{tr}(D_+^2)} + \sqrt{\operatorname{tr}(D_-^2)} \leq 2\sqrt{\operatorname{tr}(B^2)}$ and $\|D_\pm\|_\infty \leq \|B\|_\infty = \|B\|_\infty$. Plugging these relations into the bounds finishes the proof. □

Now we apply the above result to a sum of centered squares of standard normal variables weighted by the coordinates of a unit vector.

Corollary 5.2.3. *For a unit vector $\boldsymbol{u} \in I\!R^n$ and a vector of independent standard normal random variables $\boldsymbol{\varepsilon} \in I\!R^n$, it holds with $\|\boldsymbol{u}\|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} |u_i|$*

$$I\!P\left(\left|\sum_{i=1}^n u_i(\varepsilon_i^2 - 1)\right| \geq 4\mathtt{x}^{1/2} + 2\|\boldsymbol{u}\|_\infty \mathtt{x}\right) \leq 2\exp(-\mathtt{x}).$$

Proof. The result follows from Corollary 5.2.2 as $\mathrm{v}^2 = \|\boldsymbol{u}\|^2 = 1$ and $\mathtt{p} = \sum_{i=1}^n u_i$. □

The bounds so far relied on the independence of the components of the standard normal vector $\boldsymbol{\varepsilon} \in I\!R^n$. Now we will formulate results in the case of a dependency structure. Let us assume given a normal vector $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{V})$ with the covariance matrix $\mathbb{V} = (\sigma_{ij})_{1 \le i,j \le n} \in I\!R^{n \times n}$ and $\lambda_{\max}(\mathbb{V}) \le \lambda^*$. Let $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in I\!R^n$ and consider the quadratic form

$$Q = \sum_{i=1}^n u_i \xi_i^2.$$

We want to show concentration of Q around its mean. First we rewrite $\boldsymbol{\xi} = \mathbb{V}^{1/2}\boldsymbol{\varepsilon}$ as a transformation of a standard normal vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{1}_n)$. This brings us back to the framework of Corollary 5.2.3. We write $B = \mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}^{1/2}$ with the matrix $\boldsymbol{U} \stackrel{\text{def}}{=} \operatorname{diag}(\boldsymbol{u})$, which gives

$$Q = \sum_{i=1}^n u_i \xi_i^2 = \boldsymbol{\xi}^\top \boldsymbol{U} \boldsymbol{\xi} = (\mathbb{V}^{1/2}\boldsymbol{\varepsilon}^\top)\boldsymbol{U}\mathbb{V}^{1/2}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top B \boldsymbol{\varepsilon}.$$

Therefore the bound $\|\mathbb{V}\|_{\mathrm{op}} \le \lambda^*$ implies

$$\begin{aligned} \lambda &= \lambda(B) = \|\mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}^{1/2}\|_{\mathrm{op}} \le \lambda^* \, \|\boldsymbol{u}\|_\infty \,, \\ \mathrm{v}^2 &= \operatorname{tr}(B^2) = \operatorname{tr}\bigl(\mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}\boldsymbol{U}\mathbb{V}^{1/2}\bigr) \le \lambda^* \operatorname{tr}\bigl(\boldsymbol{U}\mathbb{V}\boldsymbol{U}\bigr) \le \lambda^{*2} \|\boldsymbol{u}\|^2. \end{aligned}$$

This gives us a handy reformulation of the concentration results in the case, where we want to neglect a certain part of the dependency structure in a quadratic form.

Corollary 5.2.4. *For any vector $\boldsymbol{u} \in I\!R^n$, and a normal zero mean vector $\boldsymbol{\xi} \sim \mathrm{N}(0, \mathbb{V})$ in $I\!R^n$ with $\|\mathbb{V}\|_{\mathrm{op}} \le \lambda^*$, it holds*

$$I\!P\biggl(\Bigl| \sum_{i=1}^n u_i (\xi_i^2 - I\!E\xi_i^2) \Bigr| \ge 4\lambda^* \, \|u\| \mathtt{x}^{1/2} + 2\lambda^* \, \|\boldsymbol{u}\|_\infty \mathtt{x} \biggr) \le 2\exp(-\mathtt{x}).$$

and if $\|\boldsymbol{u}\| = 1$:

$$I\!P\biggl(\Bigl| \sum_{i=1}^n u_i (\xi_i^2 - I\!E\xi_i^2) \Bigr| \ge 4\lambda^* \, \mathtt{x}^{1/2} + 2\lambda^* \, \|\boldsymbol{u}\|_\infty \mathtt{x} \biggr) \le 2\exp(-\mathtt{x}).$$

The results can be extended to the case of non-diagonal $\boldsymbol{U}$. In the case of a unit vector $\|\boldsymbol{u}\| = 1$, we can bound $\|\boldsymbol{u}\|_\infty \le 1$, which eliminates the

dependency on $\boldsymbol{u}$ completely. It is worth noting that the identity $\|\boldsymbol{u}\| = 1$ implies $\|\boldsymbol{u}\|_\infty \leq 1$. In most of the situations we are considering, the supremum norm $\|\boldsymbol{u}\|_\infty$ will be very small compared to $\|\boldsymbol{u}\|$, meaning that the $\sqrt{\mathtt{x}}$-term will normally dominate the bound, which means that we can neglect the term with $\|\boldsymbol{u}\|_\infty$.

5.3 Bounds on the total variation distance between two Gaussian vectors

In this section, we specify a bound on the total variation distance between two Gaussian measures following standard arguments based on Pinsker's inequality (Lemma 2.5(i), [Tsybakov, 2008]) and a bound on the Kullback-Leibler divergence. We recall the definition of the Kullback-Leiber divergence $\mathcal{K}(I\!P_0, I\!P_1)$ of two probability measures $I\!P_0$ and $I\!P_1$.

$$\mathcal{K}(I\!P_0, I\!P_1) \stackrel{\text{def}}{=} -I\!E_0 \log \frac{dI\!P_1}{dI\!P_0},$$

where $I\!E_0$ denotes the expectation associated with $I\!P_0$. We assume given two p-dimensional centered normal vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$ and $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, S^\flat)$ with covariance matrices $S, S^\flat$. Let $T : I\!R^p \to I\!R^{|E|}$ be Borel-measurable, E being some countable index set, and $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^\flat)$. Now we want to bound the total variation distance between the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the following conditions: There exist $\delta, \Delta \geq 0$, such that

$$\|S^{-1/2} S^\flat S^{-1/2} - \mathbf{1}_p\|_{\text{op}} \leq \delta \leq 1/2, \tag{5.3.1}$$

$$\|S^{-1/2} S^\flat S^{-1/2} - \mathbf{1}_p\|_{\text{Fr}} \leq \Delta. \tag{5.3.2}$$

The next lemma gives an application of Pinsker's inequality to our case of two multivariate normals. This proof follows the same lines as the one of Lemma A.7 in [Spokoiny and Zhilova, 2014].

Lemma 5.3.1. *Let $I\!P_0 = \mathcal{N}(0, S)$ and $I\!P_1 = \mathcal{N}(0, S^\flat)$ for some invertible matrices $S, S^\flat \in I\!R^{p \times p}$ and let* (5.3.1) *and* (5.3.2) *be satisfied, then for any given Borel measurable set $A \subset I\!R^p$, it holds*

$$\big|I\!P_0(A) - I\!P_1(A)\big| \leq \frac{1}{2}\Delta^2.$$

Proof. By the change of variables $\boldsymbol{u} = S^{-1/2}\boldsymbol{x}$ we reduce the problem to the case when $I\!P_0$ is standard normal in $I\!R^p$, while $I\!P_1 = \mathcal{N}(0, B)$ with $B \stackrel{\text{def}}{=} S^{-1/2} S^\flat S^{-1/2}$

$$2 \log \frac{dI\!P_1}{dI\!P_0}(\boldsymbol{\gamma}) = \log \det(B) - \boldsymbol{\gamma}^\top B \boldsymbol{\gamma} + \|\boldsymbol{\gamma}\|^2$$

with $\boldsymbol{\gamma}$ standard normal and

$$\mathcal{K}(I\!P_0, I\!P_1) = -I\!E_0 \log \frac{dI\!P_1}{dI\!P_0} = -0.5 \log \det(B) + 0.5 \operatorname{tr}(B - \mathbf{1}_p).$$

Let λ_j be the j th eigenvalue of $B - \mathbf{1}_p$. The condition $\|B - \mathbf{1}_p\|_{\text{op}} \leq 1/2$ implies $|\lambda_j| \leq 1/2$ and $\|S^{-1/2} S^\flat S^{-1/2} - \mathbf{1}_p\|_{\text{Fr}} \leq \Delta$ implies $\operatorname{tr}\left((B - \mathbf{1}_p)^2\right) \leq \Delta^2$. Therefore by the inequality: $a - \log(1 + a) \leq a^2$ for $a \leq 0.5$:

$$\begin{aligned}
\mathcal{K}(I\!P_0, I\!P_1) &= 0.5 \sum_{j=1}^{p} (\lambda_j - \log(1 + \lambda_j)) \\
&\leq 0.5 \sum_{j=1}^{p} \lambda_j^2 \\
&\leq 0.5 \operatorname{tr}(B - \mathbf{1}_p)^2 \leq 0.5 \Delta^2.
\end{aligned}$$

It follows by Pinsker's inequality (Lemma 2.5(i), [Tsybakov, 2008]) that

$$\sup_{A \in \mathcal{B}_p} |I\!P_0(A) - I\!P_1(A)| \leq \sqrt{\frac{1}{2} \mathcal{K}(I\!P_0, I\!P_1)} \leq \frac{1}{2} \Delta^2.$$

□

As stated at the beginning of section 5.1, we can get a bound on the Frobenius norm in terms of the operator norm and vice versa. We recall that for a matrix $M \in \mathbb{R}^{p \times p}$, we get the bound

$$\|M\|_{\text{Fr}} \leq p \|M\|_{\text{op}}.$$

This is a good bound in the case, when the eigenvalues of the matrix are almost all of the same magnitude. On the other hand, the operator norm is bounded by the Frobenius norm which means that we can state our results entirely in terms of one of the two quantities by incurring a possible loss in sharpness of the bounds. To make these general results easy to use for our

purposes, we formulate a corollary, which adapts the lemma to the type of sets we are considering mainly in our results — threshold bounds of a set of measurable transformations.

Corollary 5.3.2. *Let two p-dimensional centered normal vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$, $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, S^\flat)$ with covariance matrices $S, S^\flat$ be given and assume* (5.3.1) *and* (5.3.2). *Then for any measurable mapping $T\colon \mathbb{R}^p \to \mathbb{R}^{|E|}$ with E being a countable set, and any set of values $(q_\eta)_{\eta\in E} \subset \mathbb{R}^+$, the random vectors $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^\flat)$ satisfy*

$$\left|\mathbb{P}\big(\max_{\eta\in E}(X_\eta - q_\eta) > 0\big) - \mathbb{P}\big(\max_{\eta\in E}(Y_\eta - q_\eta) > 0\big)\right| \le \Delta/2.$$

Proof. This is a simple application of Lemma 5.3.1 with $E = \bigcup_{\eta\in E}\{\boldsymbol{x} \in \mathbb{R}^p\colon T_\eta(\boldsymbol{x}) > q_\eta\}$. □

Chapter 6

Conclusions & Outlook

In this thesis, we have introduced a versatile Lepski-type method for doing model selection without knowledge of the noise structure. The theoretical properties seem convincing and the simulation results also look promising. The method can treat a whole array of different estimation problems in ordered model selection in a unified framework — as well from the point of the theoretical analysis as from the point of view of the implementation of the algorithm. The method tunes its critical values to account for the dependencies between the different models and gives sharper multiplicity correction than a simple Bonferroni bound. The assumption on the minimal smoothness of the true function in the boostrap-setup is a weak one (just assuming a Hölder smoothness larger than $1/4$) and the critical dimension of the maximal model dimension p, which is of order $\sqrt{n/\log(n)}$ does not restrict the method too much. The theoretical results and the simulations both show that the method is robust against the choice of the calibration dimension of the presmoothing estimators. This seems important, as it makes the method usable in practice. An obvious interesting idea for further study would be to adapt the arguments to a general non-Gaussian case. While without bias, the stochastic part in our arguments could probably be controlled by arguments following [Spokoiny and Zhilova, 2014], with a significant bias for some of the models, it less clear how to accomplish a sensible extension to a general non-Gaussian case.
Another possibility for future work lies in the design of a faster version of

the algorithm and its theoretical study. While the boostrap-algorithm works well, it is still quite slow due to the fact that one needs to recalibrate for each data set $\boldsymbol{Y}$ and estimation target W. One way to make the method faster, would be to leave out some of the comparisons in the acceptance condition. One would need to study to what point the theoretical properties we have shown for the full-comparisons algorithm can be transported to a faster pruned alternative.

Bibliography

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE*, 19(6):716 – 723.

[Arlot, 2009] Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Statist.*, 3:557–624.

[Arlot and Bach, 2009] Arlot, S. and Bach, F. R. (2009). Data-driven calibration of linear estimators with minimal penalties. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 46–54. Curran Associates, Inc.

[Arlot and Celisse, 2010] Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79.

[Barron et al., 1999] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

[Bauer and Reiss, 2008] Bauer, F. and Reiss, M. (2008). Regularization independent of the noise level: an analysis of quasi-optimality. *Inverse Problems*, 24(5):055009.

[Beran, 1986] Beran, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1295–1298.

[Birgé, 2001] Birgé, L. (2001). *An alternative point of view on Lepski's method*, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 113–133. Institute of Mathematical Statistics, Beachwood, OH.

[Birgé and Massart, 2007] Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

[Cai and Low, 2003] Cai, T. T. and Low, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.*, 31(4):1140–1153.

[Cai and Low, 2005] Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *Ann. Statist.*, 33(5):2311–2343.

[Cavalier et al., 2002] Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874.

[Cavalier and Golubev, 2006] Cavalier, L. and Golubev, Y. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.*, 34(4):1653–1677.

[Chernozhukov et al., 2014] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.*, 42(5):1787–1818.

[Dalalyan and Salmon, 2012] Dalalyan, A. S. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355.

[Giné and Nickl, 2010] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.

[Hanson and Wright, 1971] Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083.

[Hsu et al., 2012] Hsu, D., Kakade, S. M., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

[Lepski, 1990] Lepski, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

[Lepski, 1991] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659.

[Lepski, 1992] Lepski, O. V. (1992). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 37(3):468–481.

[Lepski et al., 1997] Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25(3):929–947.

[Lepski and Spokoiny, 1997] Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6):2512–2546.

[Mallows, 1973] Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15(4):661–675.

[Mammen, 1993] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, 21(1):255–285.

[Massart, 2007] Massart, P. (2007). *Concentration inequalities and model selection.* Number 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer.

[Rudelson and Vershynin, 2013] Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18(82):1–9.

[Shao, 1997] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, (7):221–264.

[Spokoiny and Vial, 2009] Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37:2783–2807.

[Spokoiny and Zhilova, 2013] Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113.

[Spokoiny and Zhilova, 2014] Spokoiny, V. and Zhilova, M. (2014). Bootstrap confidence sets under model misspecification. *ArXiv e-prints*.

[Stein, 1981] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151.

[Tropp, 2015] Tropp, J. A. (2015). An Introduction to Matrix Concentration Inequalities. *to appear in Found. Trends Mach. Learning*.

[Tsybakov, 2000] Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l'Academie des Sciences - Series I - Mathematics*, 330(9):835 – 840.

[Tsybakov, 2008] Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer.

[Willrich and Spokoiny, 2015] Willrich, N. and Spokoiny, V. (2015). Bootstrap tuning in ordered model selection. (manuscript to be submitted).

[Wu, 1986] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295.

[Yang, 2005] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950.